EVALUATING THE SECURITY OF ANONYMIZED BIG GRAPH/STRUCTURAL DATA

A Dissertation
Presented to
The Academic Faculty

by

Shouling Ji

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
December 2015

EVALUATING THE SECURITY OF ANONYMIZED BIG GRAPH/STRUCTURAL DATA

Approved by:

Professor Matthieu Bloch,
Committee Chair
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Raheem Beyah, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Faramarz Fekri
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor John Copeland
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Mustaque Ahamad
School of Computer Science
Georgia Institute of Technology

Date Approved: 11 November 2015

This dissertation is dedicated to my family.

ACKNOWLEDGEMENTS

I would like to show my great gratitude to all of those people who supported and helped me to complete my dissertation. Their generous help made this dissertation possible.

First of all, I am deeply grateful to my advisor, Dr. Raheem Beyah, for his inspiration, guidance, thoughts, and friendship. Dr. Beyah always gives me the greatest tolerance and support during my Ph.D. studies. The discussion with him is thought-provoking and the source of my ideas and inspirations. Dr. Beyah is the perfect model for me both in research and in life.

I would like to thank my committee members, Dr. Matthieu Bloch, Dr. Faramarz Fekri, Dr. John Copeland, and Dr. Mustaque Ahamad. Thank them very much for spending time to serve my Ph.D. committee and for their valuable comments to improve this dissertation and my research. As the committee chair, Dr. Bloch spent a lot of efforts to arrange my proposal exam and final defense. Dr. Fekri and Dr. Copeland also gave me a lot of help when I was in their classes.

Many thanks go to my first Ph.D. advisors Dr. Yingshu Li and Dr. Zhipeng Cai. By their broad academic vision and incisive academic perspectives, they provided me many suggestions on my research with foresight and sagacity. In addition to research, they gave me a lot of help on my life and career planning.

Many thanks to Dr. Prateek Mittal, Dr. Xin Hu, Dr. Ting Wang, Dr. Mudhakar Srivatsa, Dr. Marc Stoecklin, Dr. Josyula Rao, Dr. Jiyong Jang, Dr. Dhilung Kirat, Dr. Douglas L. Schales, and Dr. Neil Zhenqiang Gong. Dr. Mittal gave me a lot of constructive advice to conduct data privacy research. Dr. Hu, Dr. Wang, and Dr. Srivatsa helped me a lot when I was at IBM T. J. Watson Research Center and

provided me many valuable resources. Dr. Stoecklin, Dr. Rao, Dr. Jang, Dr. Kirat, and Dr. Schales also gave me plenty of help when I was at IBM Research. Dr. Gong provided me many suggestions and data resources to conduct security and privacy research.

I would like to express my appreciation to Prof. Jinbao Li and Prof. Jianzhong Li, who brought me to the research community and encouraged me to make progress in the research world.

Special thanks go to my group members, fellow students, friends, and colleagues, Weiqing Li, Shukun Yang, Qinchen Gu, Changchang Liu, Wei-Han Lee, Zhigong Li, Dr. Selcuk Uluagac, Xiaojing Liao, David Formby, Dr. Tielei Wang, Dr. Jing (Selena) He, Dr. Mingyuan Yan, Yueming Duan, Meng Han, Guoliang Liu, Chenguang Kong, Xuhong Zhang, Yumei Lu, Sha Liu, Dr. Wenzhuo Wu, Dr. Nana Li, Sang Shin Jung, Samuel Litchfield, Dr. Troy Nunnally, Shruthi Ravichandran, and Dr. Marco Valero. They provided me invaluable suggestions and help for my study and research. Especially, without of the help from Weiqing, Shukun, and Qinchen, I cannot finish this dissertation as expected. Thank all of my friends.

Last but not least, I would like to thank my family for their continuous support, understanding, and help. They are there whenever I need them. This dissertation is dedicated to them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

We studied the security of anonymized big graph data. Our main contributions include: *new De-Anonymization (DA) attacks*, *comprehensive anonymity, utility, and de-anonymizability quantifications*, and *a secure graph data publishing/sharing system SecGraph*.

New DA Attacks. We present two novel graph DA frameworks: *cold start single-phase Optimization-based DA* (ODA) and *De-anonymizing Social-Attribute Graphs* (De-SAG). Unlike existing seed-based DA attacks, ODA does not priori knowledge. In addition, ODA's DA results can facilitate existing DA attacks by providing more seed information. De-SAG is the first attack that takes into account both graph structure and attribute information. Through extensive evaluations leveraging real world graph data, we validated the performance of both ODA and De-SAG.

Graph Anonymity, Utility, and De-anonymizability Quantifications. We developed new techniques that enable comprehensive graph data anonymity, utility, and de-anonymizability evaluation. First, we proposed the first seed-free graph de-anonymizability quantification framework under a general data model which provides the theoretical foundation for seed-free SDA attacks. Second, we conducted the first seed-based quantification on the perfect and partial de-anonymizability of graph data. Our quantification closes the gap between seed-based DA practice and theory. Third, we conducted the first attribute-based anonymity analysis for Social-Attribute Graph (SAG) data. Our attribute-based anonymity analysis together with existing structure-based de-anonymizability quantifications provide data owners and researchers a more complete understanding of the privacy of graph data. Fourth, we conducted the first graph Anonymity-Utility-De-anonymity (AUD) correlation quantification and

provided close-forms to explicitly demonstrate such correlation. Finally, based on our quantifications, we conducted large-scale evaluations leveraging 100+ real world graph datasets generated by various computer systems and services. Using the evaluations, we demonstrated the datasets' anonymity, utility, and de-anonymizability, as well as the significance and validity of our quantifications.

SecGraph. We designed, implemented, and evaluated the first uniform and open-source Secure Graph data publishing/sharing (SecGraph) system. SecGraph enables data owners and researchers to conduct accurate comparative studies of anonymization/DA techniques, and to comprehensively understand the resistance/vulnerability of existing or newly developed anonymization techniques, the effectiveness of existing or newly developed DA attacks, and graph and application utilities of anonymized data.

CHAPTER I

INTRODUCTION

With the rapid development of information technology, a huge amount of data are generated from various computer systems [30, 38, 53, 88, 97, 122, 129, 131, 137, 142, 152, 158, 159, 160]. Data that can be modeled by *graphs* as shown in Fig.1, where *nodes* represent *data items* (e.g., users of social networks) and *edges/links* represent the *relationships* (e.g., the friendships among Facebook users, the contact relations among email users) among data items, are considered *graph/structural/complex data* (for convenience, we use *graph data* in the rest of this dissertation as done in relevant works [1, 27, 61, 62, 63, 104, 127, 152, 159]).

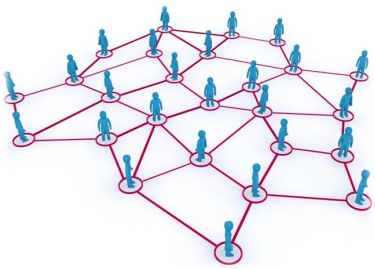


Figure 1: Graph data.

1.1 Overview of Graph Data

Nowadays, many computer systems generate graph data [152, 159]. Below, we summarize representative computer-generated graph data.

- *Social Network Data.* It is natural to represent social networks (e.g., Facebook [46], Google+ [51], Twitter [133], LinkedIn [87], YouTube [153], LiveJournal

[89], Orkut [110], Slashdot [126], and Pokec [116]) as graphs, where nodes denote users and links/edges denote the social relationships (friendship, circle-relationship, follow relationship, etc.) among users;

- *Communication Data.* Another typical category of graph data is communication data, including phone-call networks [71, 73, 100, 109], email networks [67, 80, 81], wiki-Talk networks [78, 79], instant message networks [127], etc. To represent communication networks, users are modeled by nodes and the communication relationship (phone calls, emails, talks, etc.) are modeled by links ;

- *Mobility Traces.* Mobility traces, e.g., WiFi traces [127], Bluetooth traces [127], check-ins [115], usually consist of records of format (*user ID, latitude, longitude, timestamp, location ID*). They can be transferred to user-connect graphs by applying sophisticated data processing techniques (e.g., entropy-based techniques) [115, 127], where nodes represent users and links/edges indicate the co-appearance or connection relation;

- *Epidemiological and Health-care Data.* A large amount of healthcare data is in graph form, leveraging which health-care professionals can study disease propagation as well as other social health problems [22, 28, 58]. For instance, to study the sexual contact-based disease transmission, an adolescent romantic and sexual network is published in [28], which consists of a population of over 800 adolescents residing in a mid-sized town in the mid-western United States.

- *Collaboration Networks.* Collaboration networks, e.g., Arxiv [24, 80], the computer science collaboration network DBLP and ArnetMiner [2, 3], represent the collaboration relationships among researchers. Straightforwardly, collaboration networks can be modeled by graphs where nodes represent researchers and links represent collaborations.

- *Citation Networks* [48, 60]. Citation networks carry the citation information among research papers, which are naturally graph data.

- *Web Graphs* [21, 25, 81]. Web graphs indicated the link information among web pages, where nodes represent web pages and edges represent hyperlinks among them.

- *Internet Peer-to-Peer Networks and Other Network Topologies* [80, 121]. Peer-to-Peer networks and other network topologies can be modeled by graph data, where nodes represent network terminals in the network and edges represent the connections among them.

- *Autonomous System Graphs* [60]. The graph of routers comprising the Internet can be organized into sub-graphs called Autonomous Systems (AS) [60]. Each AS exchanges traffic flows with some neighbors (peers). Therefore, graphs can be constructed to represent *who-talks-to-whom* relationships among AS.

1.2 Need for Graph Data Sharing

Graph data sharing has important implications for research, government, commercial, and healthcare applications. Below, we discuss some typical graph data sharing scenarios.

1.2.1 Academic Research

As it has been well known, real-world data publishing/sharing/transferring is the most valuable resource for academic research, e.g., personalized advertising, sense/decision-making, influence maximization, innovation/disease diffusion, similar users searching, user classification, reliable email, secure routing, and Sybil detection [4, 5, 27, 61, 63, 104, 127, 152, 159]. In this subsection, we focus on the scenarios of publishing/sharing/transferring graph data to academia for research.

During the annual KDD Cup events, several datasets (including graph datasets) are published for data mining and knowledge discovery tasks [4]. For instance, several network topological structure datasets, social network datasets, customer relationship graphs are published or shared with researchers. Similarly, many other academic events/instituions regularly provide graph data to the research community [5, 6, 102, 103, 104]. Recently, Twitter introduced its data sharing project to the academic community, named *Twitter Data Grants*, through which Twitter's public and historical data are accessible [6].

> *Today were introducing a pilot project were calling Twitter Data Grants, through which well give a handful of research institutions access to our public and historical data.*
>
> *With more than 500 million Tweets a day, Twitter has an expansive set of data from which we can glean insights and learn about a variety of topics, from health-related information such as when and where the flu may hit to global events like ringing in the new year. To date, it has been challenging for researchers outside the company who are tackling big questions to collaborate with us to access our public, historical data. Our Data Grants program aims to change that by connecting research institutions and academics with the data they need.*

In order to promote real-world data driven research, many other real-world graph data have been shared with researchers, e.g., Facebook data [5, 7], QQ data [8], Microblog data [5], Citation data [5].

1.2.2 Government Data Mining Tasks

In addition to being leveraged by academia for research, graph data are frequently shared/transferred for government data mining tasks. For instance, customer understanding and international fraud detection can be achieved by leveraging the structure

and pattern analysis of phone-call networks [139]. Furthermore, communication data (e.g., phone-call networks, email networks) can also be applied to serious national security data mining tasks, such as graph theory-based terrorist analysis [55]. Recently, it has been shown that a lot that government agencies employ graph data (e.g., phone data of Verizon, social graph data Google and Facebook, email networks, instant messaging networks, and video conference data) for several kinds of data mining tasks (e.g., *fighting terrorism*) [40].

> *In the name of fighting terrorism, the US government has been mining data collected from phone companies such as Verizon for the past seven years and from Google, Facebook, and other social media firms for at least four years, according to government documents leaked this week to news organizations.*
>
> *The two surveillance programs one that collects detailed records of telephone calls, the other that collects data on Internet-based activities such as e-mail, instant messaging, and video conferencing were publicly revealed in "top secret" documents leaked to the British newspaper the Guardian and the Washington Post. Both are run by the National Security Agency (NSA), the papers reported.*

In addition, some companies have been reported to sell graph data-based data mining solutions to governments [40].

1.2.3 Business Applications

Data sharing/transferring is a standard practice for companies. For instance, as described in their privacy policies [9, 10, 11], Google, Facebook, and Twitter share their data with business partners for personalized advising, under which cost savings and maximized advertising effectiveness can be achieved. In addition to advertising, graph data are also shared among companies to build enterprise applications to improve

business decisions. For instance, recently, Twitter and IBM announced a significant partnership that will involve Twitter sharing its data with IBM for integration into IBMs enterprise solutions, including the Watson cloud platform [114].

> *Twitter and IBM announced a significant partnership today that will involve Twitter sharing its data with IBM for integration into IBMs enterprise solutions, including the Watson cloud platform. The deal means IBM will gain access to the Twitter firehose, allowing businesses to incorporate insights gained from the social network into their decision-making processes.*
>
> *Additionally, the two companies will also be teaming up to build a unique collection of enterprise solutions, they say, which puts IBM into a different category than some of Twitters other data partners, who generally just ingest the data for use in their own systems.*
>
> *IBM says the companies will collaborate to build enterprise applications to improve business decisions across industries and professions, beginning with applications and services for sales, marketing and customer service. They will also work together on industry-specific solutions, including those for banking, consumer products, transportation and retail* [114].

We also examine the privacy policies of some companies as follows.

> *According to google, information used during registration, used while using the services are collected by google and might be given to trusted parties for processing based on googles instructions and in compliance with googles Privacy Policy. The information can be given to third parties with the users consent and given to the users domain administrators as well as for legal reasons. The aggregated data with no personally identifiable information can be released publicly and with googles partners.*

–Google Plus Privacy Policy [9].

Facebook have information on users from registration as well as from posts a user makes or shares. According to the policy, personal information can be edited during use and as for the meta information in the things shared, that is up to the user to remove. In addition with the user consent Facebook can have the contact information from the users email account, transaction information from purchases on the site, as well location information. In addition to information specifically shared with Facebook, they also have information from third party such as ad companies, information from other users, cookies, and user devices. With the information collected, Facebook can contact the user, send targeted advertisement to the user, improve their service, and make suggestions to user. With the information Facebook have, they can share the payment information to complete a purchase, send email to invite others on behalf on the user, share information with marketers, help others to find you, and give search engines access to your public information.

–Facebook Privacy Policy [10].

Twitter have information that are provided during account creation as well as other information the user choose to give such as the users phone number for SMS fraud protection, picture, and location. Most of the information on twitter are publicly available such as lists created, the people a user follow and the people that follows a user. Twitter can publish a user location information and phone numbers with the users consent and this includes the location information from IP, user device, as well as from cell towers. In addition to location information Twitter have information on how a user interacts with links on Twitter and from email by Twitter, as

well as cookie information, log data, widget data and payment information. With all the information Twitter can share them with users consent, share with service providers with uses that follows Twitters privacy information and for uses that are state by Twitter. If making purchases on Twitter, Twitter can share information such as address and name. In addition user information might be sold in case that Twitter is in a part of a buy out or merger or if the information is need for legal reasons. Lastly public information can be shared to others for reasons such as advertisers whose link a user clicked on.

–Twitter Privacy Policy [11].

From the above policies, users data (graph data) are clearly shared among partners.

1.2.4 Healthcare Applications

Graph data are also published/shared/transferred for many civil applications. A typical such scenario is to analyze the propagation of infectious diseases, e.g., the flu, HIV, Ebola [12, 13, 74]. Real-world graph data are valuable to accurate disease propagation analysis. As shown in [28], when analyzing sexual contact-based disease diffusion, real sexual networks-based analysis is very different from that leveraging simulated or randomly generated graph data. Recently, when studying the Ebola Outbreak 2014, the Ebola Hemoragic Fever propagation in a modern city is modeled and analyzed based on the social graphs and other data [13].

1.2.5 Other Scenarios

Graph data are widely available in many other scenarios.

- For conducting research, developing web and mobile applications, designing data visualizations, and other applications, government agencies regularly release

data by law [14].

- Many graph data can be crawled employing an API or screen-scraping, e.g., Google+ [15], Facebook [7, 16], Twitter [1, 16], YouTube [1, 16, 96], LinkedIn [86].

- Graph data are widely available on many data sharing websites [1, 15, 16, 17]. For instance, many social network data, communication network data, mobility traces, collaboration data, autonomous system graphs are available at Stanford SNAP [16], ASU Network Data Repository [1], Dartmouth CRAWDAD [18], UCI Network Data Repository [19], CMU Datasets [17], etc.

- Recently, with the emergence of *data brokers*, many graph data, especially the sensitive data such as medical records, financial information, credit reports, social relations, and other personal profiles, are easily obtained [29, 43, 132].

  .....

  Consumer data companies are scooping up huge amounts of consumer information about people around the world and selling it, providing marketers details about whether you're pregnant or divorced or trying to lose weight, about how rich you are and what kinds of cars you drive. But many people still don't know data brokers exist.

  .....

  As we highlighted last year, some data companies recordand then resellall kinds of information you post online, including your screen names, website addresses, interests, hometown and professional history, and how many friends or followers you have [29].

1.3 Graph Data Security and Utility

Different from traditional *relational data* (e.g., *tabular data*, *set-valued data*), where data items are structurally independent of each other, *the most notable characteristic of the data items of graph data is that they are structurally correlated with each other in addition to the semantic information they carry* [1, 27, 152, 159]. For instance, a user of a social network is correlated with other users in the network in addition to the social profiles associated with him/her. On one hand, *the correlations of graph data items enable many new applications*. On the other hand, *these correlations allow graph data to suffer security and privacy threats since adversaries can leverage them to infer private information of the users/systems who generated the graph data*. Recent research by us and others has shown that simply anonymized graph data can be successfully de-anonymized in large-scale by *Structure-based De-Anonymization (SDA) attacks* [27, 61, 63, 64, 104, 127]. The main idea of SDA attacks is to de-anonymize anonymized users in terms of their uniquely distinguishable structural characteristics.

To protect graph users' privacy when sharing graph data, several anonymization techniques have been proposed, which can be classified into six categorizes: *Naive ID Removal*, *Edge Editing (EE) based techniques* [152], *k-anonymity based techniques* [38, 88, 156, 158, 160], *Aggregation/Class/Cluster based techniques* [30, 53, 131], *Differential Privacy (DP) based techniques* [117, 118, 122, 137, 142], and *Random Walk (RW) based schemes* [97]. Basically, these anonymization techniques try to perturb the original graph structure to protect users' privacy while preserving as much data utility as possible.

Therefore, existing anonymization schemes can be evaluated from two perspectives: *data utility preservation* and *resistance to DA attacks* (we use DA and SDA interchangeably). *However, most, if not all, existing graph anonymization works*

have not been significantly evaluated in terms of their utility performance. Specifically, most existing graph anonymization works only conducted limited evaluation on their utility preservation, e.g., *degree distribution*, *path length distribution*, *cluster coefficient*, which are insufficient to understand their value to high-level network mining tasks and applications, e.g., sense/decision-making, similar users searching, user classification, reliable email, influence maximization. *More Surprisingly, although we already have many sophisticated anonymization techniques (e.g., [88, 97, 152, 158]) and powerful DA attacks (e.g., [63, 69, 104, 112]), it is still an open problem whether existing state-of-the-art anonymization techniques can defend against modern SDA attacks* (i.e., the *practical vulnerability* of anonymized datasets). This is because of the *incomplete evaluation* of existing anonymization and DA works. For anonymization works, they usually only evaluate the data utility performance of their proposed techniques (although some work did provide a theoretical security guarantee, however, these guarantees usually do not hold due to improper assumptions or incomplete considerations). For DA works, they usually evaluate their attack performance without actually applying state-of-the-art anonymization techniques (e.g., k-anonymity based schemes, DP based schemes, and RW based schemes) to test their technique.

Furthermore, recently, the concept of *graph data de-anonymizability quantification* has also garnered significant attention[61, 63, 113], where researchers study that, **based only on graph data's structural information, why graph data can be de-anonymized, what are the DA conditions, and how many users are de-anonymizable**, i.e., *graph data de-anonymizability quantification* can quantitatively examine how vulnerable/de-anonymizable any graph dataset is given its structure. Therefore, graph data de-anonymizability quantification techniques can be employed to examine the *theoretical vulnerability* of both *raw* and *anonymized graph data*, and can therefore evaluate the effectiveness of an anonymization scheme. Furthermore, the quantification results can serve as auxiliary information that is useful for future

effective anonymization technique design and DA attack evaluation. However, existing de-anonymizability quantifications are limited because *they treat all the users within a graph as structurally equivalent and overlook their structural differences.* In practice, different users may have very different structural importance, e.g., the users with the maximum and minimum degrees are structurally different. Therefore, existing quantification results are incomplete with regards to quantifying graph users' actual de-anonymizability in terms of their structural importance.

The two overarching goals of this dissertation are (*i*) *to help graph data holders understand if and how their data should be shared*; and (*ii*) *to provide graph data security/privacy researchers a uniform platform to comprehensively study, evaluate, and compare existing/newly developed graph data anonymization, utility evaluation, DA, and de-anonymizability quantification techniques.* This will be accomplished by developing (*i*) new techniques that enable comprehensive graph data anonymization-utility-de-anonymization evaluation, accurate structure-based de-anonymizability quantification, and complete utility-de-anonymizability analysis; and (*ii*) a new *practical, easily-usable, open-source and uniform* system that can systematically integrate existing and newly developed graph data anonymization, utility evaluation, DA, and utility-de-anonymizability quantification techniques.

1.4 Research Picture

We summarize the research in this dissertation in Fig.2. Specifically, we made the following contributions.

- We presented two new de-anonymization frameworks. First, following our graph de-anonymizability analysis, we proposed a novel *Optimization-based De-Anonymization* (ODA) framework. Different from existing SDA attacks (which

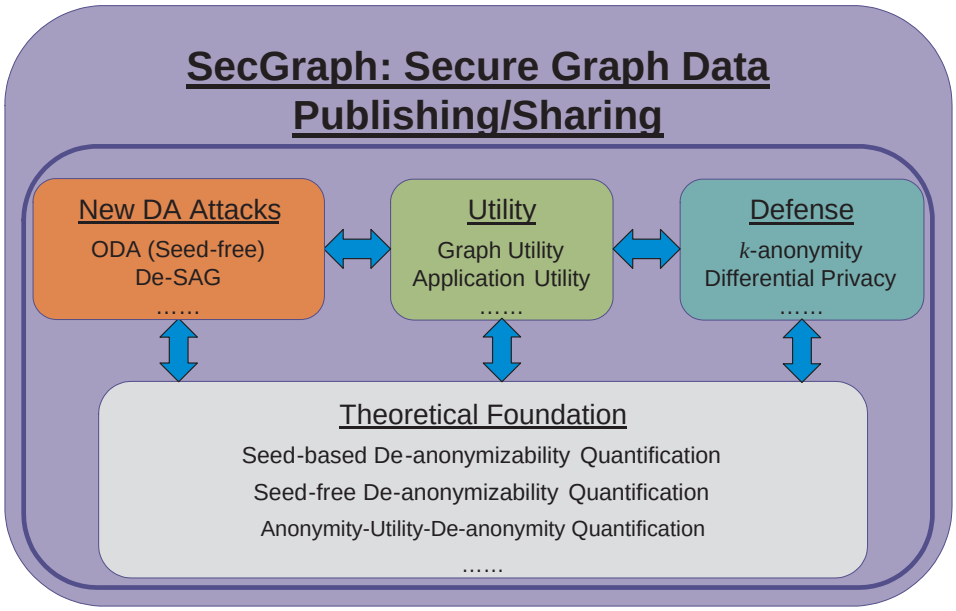


Figure 2: Research picture.

are seed-free attacks) [64, 104, 127], ODA is a *single-phase cold start* algorithm without any requirement on priori knowledge, e.g., seed/landmark mappings. We examined ODA on real graph datasets Gowalla (196,591 users and 950,327 edges) and Google+ (4,692,671 users and 90,751,480 edges). The results demonstrate that about $77.7\% - -83.3\%$ of the users in Gowalla and $86.9\% - -95.5\%$ of the users in Google+ are de-anonymizable, which implies seed-free de-anonymization is implementable and powerful in practice.

Second, according to our attribute-based anonymity analysis, we proposed a

new de-anonymization attack on graph data, namely De-anonymize Social-Attribute Graph data (De-SAG), which takes into account both graph structure and attribute information to the best of our knowledge. Through extensive evaluations leveraging real world graph data, we demonstrated that De-SAG can significantly enhance existing SDA attacks. For instance, when de-anonymizing a Facebook dataset (4,039 users, 88,234 user-user links, 1,283 attributes, 37,257 user-attribute links), De-SAG has a $3.82 \sim 10.1$ times better de-anonymization performance than state-of-the-art structure-based deanonymization attacks [63, 69].

- We developed new techniques that enable comprehensive graph data anonymity, utility, and de-anonymizability evaluation. First, we proposed the first seed-free graph de-anonymizability quantification framework under a general data model. In our quantification, we answered several fundamental open problems: why graph data can be de-anonymized based only on the topological information? what are the conditions for perfect and $(1-\epsilon)$-perfect seed-free de-anonymization, where ϵ is the error tolerated by a de-anonymization scheme? and what portion of users can be de-anonymized in a graph dataset? Thus, our quantification provides the theoretical foundation for seed-free SDA attacks.

  Second, we conducted the first seed-based quantification on the perfect and partial de-anonymizability of graph data both under the Erdös-Rényi (ER) model and in general scenarios, where the graph data can follow an arbitrary model. Therefore, our quantification can be applied to real world graph data and can quantitatively demonstrate the vulnerability of real world graph data to existing seed-based SDA attacks. Theoretically, our quantification provides the mathematical foundation for existing seed-based SDAs and closes the gap between seed-based de-anonymization practice and theory.

Third, we conducted the first attribute-based anonymity analysis for Social-Attribute Graph (SAG) data under both preliminary and general data models. By careful quantification, we explicitly demonstrate the correlation between the achievable graph anonymity and the attribute information. Our theoretical results demonstrate that the attribute information, even as non-Personal Identifiable Information (non-PII), can also lead to significant anonymity loss of graph data. Our attribute-based anonymity analysis together with existing structure-based de-anonymizability quantifications provide data owners and researchers a more complete understanding of the privacy of graph data.

Fourth, we introduced three metrics to measure the anonymity, utility, and de-anonymity of anonymized graph data, respectively. Based on these metrics, we conducted a comprehensive quantification of the correlation of graph anonymity, utility, and de-anonymity under both the mathematical ER model and a general data model. To the best of our knowledge, this is the first work on quantifying the Anonymity-Utility-De-anonymity (AUD) correlation of graph data and providing close-forms to explicitly demonstrate such correlation.

Finally, based on our quantifications, we conducted large-scale evaluations leveraging 100+ real world graph datasets generated by various computer systems and services. Using the evaluations, we demonstrated the datasets' anonymity, utility, and de-anonymizability, as well as the significance and validity of our quantifications.

- We designed and implemented a uniform and open-source Secure Graph data publishing/sharing (SecGraph) system (available at [20]). SecGraph enables

data owners to anonymize their data using state-of-the-art anonymization techniques, measure the anonymized datas graph and application utilities, and comprehensively evaluate their datas actual vulnerability against modern DA attacks. To the best of our knowledge, SecGraph is the first such system publicly available to both academia and industry. More importantly, SecGraph provides the first uniform platform that enables researchers to conduct accurate comparative studies of anonymization/DA techniques, and to comprehensively understand the resistance/vulnerability of existing or newly developed anonymization techniques, the effectiveness of existing or newly developed DA attacks, and graph and application utilities of anonymized data.

1.5 Organization

The rest of this dissertation is organized as follows: in Chapter 2, we summarize the research progress in the data anonymization and de-anonymization areas. In Chapter 3, we study the seed-free DA quantification. In Chapter 4, we study the seed-based DA quantification. In Chapter 5, we study the impact of non-PII on the anonymity of graph data. In Chapter 6, we conduct the AUD quantification for graph data. In Chapter 7, we design, implement, and evaluate SecGraph. We conclude this dissertation in Chapter 8.

CHAPTER II

RELATED WORK

In this chapter, we summarize the progress in the data anonymization and de-anonymization areas. Although we focus on graph data anonymization and de-anonymization in this dissertation, for completeness and to illustrate the evolution of techniques, we also briefly summarize the anonymization and de-anonymization schemes for non-graph data, e.g., micro/tabular data, set-valued data.

2.1 Anonymization

In this section, we summarize and classify existing anonymization techniques.

2.1.1 Micro/Tabular Data Anonymization

2.1.1.1 k-anonymity and Variants

Security/privacy is an important concern when publishing, transferring, and/or sharing data. To protect data's security and privacy, dozens of techniques have been proposed. Among them, *k-anonymity*, defined by Samarati and Sweeney [123, 128], opened a prosperous research area of data anonymization. Under k-anonymity, one user's data cannot be distinguished from at least $k-1$ other users' data in the publishing data. In general, to achieve k-anonymity is NP-hard. Therefore, many following works focus on designing efficient k-anonymization algorithms and/or extending k-anonymity to more effective *privacy models* (e.g., *ℓ-diversity* [90], *t-closeness* [83]) for specific data publishing applications.

Following [123, 128], LeFevre et al. provided a practical framework for efficient full-domain k-anonymity [77]. To improve the k-anonymity performance, Aggarwal et al. designed a $O(k)$-approximation algorithm [24] followed by Park and Shim who

further improved the approximation ratio to $O(\log k)$ [111].

To better protect users' privacy, dozens of improved privacy models of k-anonymity have been proposed. To defend against the *homogeneity attack* and *background knowledge attack* to k-anonymity, Machanavajjhala et al. proposed ℓ-*diversity* in [90], under which each equivalence class has at least ℓ well-represented values for each sensitive attribute. For protecting both identification information and sensitive relationship information in a dataset, Wong et al. extended k-anonymity to (α, k)-*anonymity* [141]. Since privacy disclosure may also happen under ℓ-diversity based on the attribute distribution, Li et al. proposed t-*closeness* in [83], which requires that the distribution of a sensitive attribute in any equivalence class should be close to the attribute distribution in the overall dataset. Similar to ℓ-diversity, to defend against the background knowledge attack on k-anonymity, Martin et al. proposed (c, k)-*safety*, where k characterizes the background knowledge and c indicates the desired privacy level [93]. To improve the accuracy of generalization based k-anonymity/ℓ-diversity, *permutation-based anonymization* was designed in [143] by Xiao and Tao and [157] by Zhang et al.

In [136], Wang and Fung proposed (X, Y)-*privacy* (including (X, Y)-*anonymity* and (X, Y)-*linkability*) to protect the privacy of sequential data releases, where X and Y are two attribute sets over the join of two sequential datasets. To address the inappropriateness of k-anonymity/ℓ-diversity in some situations, Nergiz et al. presented δ-*presence* under which an adversary cannot identify any individual as being in a dataset with certainty greater than δ [105]. To address the privacy leakage of dynamic datasets, Xiao and Tao proposed a new privacy model named m-*invariance*, where m measures the number of different users and sensitive values of each quasi-identification group [144, 145]. Considering the specific features of healthcare data, Mohammed proposed LKC-*privacy*, where L characterizes the adversary's power, and K and C measure the privacy thresholds of identity and attribute linkage, respectively

[99].

Considering that many privacy models (e.g., t-closeness) require that groups of sensitive attributes follow specified distributions, Koudas et al. proposed *P-private generation*, under which a group of sensitive attribute values can be transformed to a certain target distribution P with minimal data distortion [70]. To defend the *structure-based attack* and *label-based attack* to recommendation data, Chang et al. extended k-anonymity to a *predictive anonymization* model, where privacy, utility, and performance are considered simultaneously [34]. In [23, 91], Aggarwal et al. and Mahmood et al. generalized k-anonymity to *k-Anonymous Cluster* (k-AC), which allows more information being published without compromising privacy. Considering different personal levels of desired privacy, Choromanski relaxed k-anonymity to *b-matching* from adaptive anonymity (b is short for *bipartite graph*) [39].

k-anonymity + Utility. To make the anonymized data useful, utility-based anonymization techniques are also extensively studied. In [76], LeFevre extended k-anonymity and ℓ-diversity to *workload-aware anonymization*. In [146], Xu et al. designed two heuristic local recordings for utility-based anonymization. Similarly, Kifer and Gehrke investigated utility preserved anonymization schemes which maintain the same privacy guarantees of k-anonymity and ℓ-diversity [66]. In [32], Brickell and Shmatikov evaluated the tradeoff between privacy and utility. Their results demonstrated that even modest privacy gains require almost complete destruction of the data mining utility.

2.1.1.2 Differential Privacy

Besides k-anonymity and its variants, *Differential Privacy* (DP), introduced by Dwork [44, 45], is another popular anonymization technique to provide a provable strong privacy guarantee. Initially, DP is designed for statistical databases aiming at maximizing the accuracy of queries while minimizing the chance of privacy leakage

[44]. Following [44], many enhanced DP techniques have been proposed for different application scenarios.

In [54], Hay et al. proposed an approach to improve the accuracy of differentially private algorithms for both unattributed and universal histograms. In [98], Mohammed studied how to guarantee ϵ-DP under the *non-interactive* setting by probabilistically generalizing the raw data and then adding noise. To achieve ϵ-DP and meanwhile improve data's utility, Kellaris and Papadopoulos proposed a practical DP framework via *grouping* and *smoothing* [65]. To improve the accuracy of queries, Li et al. presented a two-stage , data and workload aware mechanism for answering sets of range queries under DP [82]. In [119], Qardaji et al. considered the scenario of differentially private releasing of *marginal contingency tables*. They introduced PriView, which computes marginal tables for a number of sets of attributes, and then reconstruct any designed k-way marginal based on these sets of attributes.

Similar to k-anonymity, many variants of ϵ-DP have been designed to better meet the privacy requirements of specific applications. In [45], Dwork et al. proposed a relaxed version of ϵ-DP, named (ϵ, δ)-*DP*, that permits both an additive term (quantified by δ) and a multiplicative term (indicated by ϵ). In [94], McSherry and Mironov applied (ϵ, δ)-DP to differentially private recommender systems. They designed and analyzed a recommender system built to provide modern privacy guarantees. In [75], Lee and Clifton et al. presented an alternative of ϵ-DP called ρ-*Differential Identifiability* (ρ-DI), which provides the same guarantees as DP while bounds the probability of individual identification by ρ. Li et al. proposed a general privacy model $(\mathbb{D}, \gamma)$-*membership privacy*, where $\mathbb{D}$ captures all states of prior knowledge of an adversary and γ limits the increase in confidence of accurate membership assertion [85]. In [84], Li et al. studied the correlation between k-anonymity and DP. They demonstrated that k-anonymization, when done "safely" and preceded with a random sampling step, meets (ϵ, δ)-DP with reasonable parameters.

2.1.2 Set-valued Data Anonymization

Different from traditional micro/tabular data, *set-valued data*, e.g., transaction data, web search queries, click streams, and transit data, refer to the data in which each record owner is associated with a set of items [35, 57]. In [57], He and Naughton extended k-anonymity to anonymize set-valued data through top-down and local generalization. Similarly, Xue et al. generalized k-anonymity and ℓ-diversity to set-valued data by nonreciprocal recording [149]. In [130], Terrovitis et al. proposed k^m-*anonymization*, which prevents an adversary from distinguishing a transaction from k transactions given him the knowledge of at most m items. In [147], Xu et al. proposed (h, k, p)-*coherence* for anonymizing transaction databases, which ensures that for an adversary of power p, the probability of identifying a transaction is limited to $1/k$ and the probability of linking an individual to a private item is limited to h. Another anonymization model is ρ-*uncertainty*, proposed by Cao et al. [33], which defends against sensitive associations without constraining the nature of an adversary's knowledge or falsifying data. Similar to for tabular data, DP is also extended to set-valued data anonymization. In [35, 36, 37], Chen et al. proposed several anonymization techniques with DP guarantee for set-valued data in different scenarios.

2.1.3 Graph Data Anonymization

Now, we discuss our main focus of this section: *anonymization techniques for graph data*. With the emergence of many graph data, e.g., social networks, Internet, WWW, collaboration networks, anonymous systems, mobility traces (which can modeled by graph data by applying sophisticated techniques [63, 64, 115, 127]), and email networks, the security and privacy issues raised during the publishing of these data have attracted a lot of attention as of recent [152, 159]. Compared to traditional relational

data (e.g., micro/tabular/set-valued data), anonymizing graph data is more challenging. First and intuitively, the structure of graph data is much more complicated. Consequently, in addition to the semantic information carried by data, the correlation and structure information among users should also be protected. Second, it is more difficult to model the auxiliary information available to adversaries, e.g., the widely available and accessible social information make the secure publishing of social data extremely challengeable [104]. Last but not least, it is more challenging to quantitatively measure the information of anonymizing graph data than relational data [159]. Therefore, anonymization techniques for relational data (micro/tabular/set-valued data) cannot be applied to graph data, and thus researchers have spent a lot of efforts to design effective graph data anonymization techniques [159]. Below, we summarize and categorize existing graph data anonymization techniques (a brief survey on graph data anonymization techniques proposed before 2008 can be found in [159]).

2.1.3.1 Naive ID Removal

To publish graph data, a straightforward method is by *naive ID removal*. Although this method has been demonstrated to be extremely vulnerable to *Structure-based De-Anonymization (SDA) attacks* (see Section 2.2), it is still widely used because of its simplicity, easy applicability, and scalability (e.g., a recent privacy leakage incident of the data indicating the locations of New York City's taxi drivers due to the poor data anonymization [50]) [27, 63, 104, 127].

2.1.3.2 Edge Editing based Anonymization

To protect graph data's privacy, Ying and Wu proposed spectrum preserved Edge Editing (EE) based schemes *Add/Del* and *Switch* [152]. Let $G(V, E)$ be a graph

dataset[1], where $V = \{i|i \text{ is a user}\}$ is the set of users and $E = \{e_{i,j}|i, j \in V$, there is a relationship between i and $j\}$ is the set of all the possible relationships (e.g., friendships, contacts, and collaboration relationships) among the users in V. Under Add/Del, k randomly chosen edges will be added to E followed by another k randomly chosen edges will be deleted from E. Under Switch, k *edge switches* are conducted, where for each edge switch, two existing edges $e_{i,j}$ and $e_{u,v}$, such that $e_{i,j}, e_{u,v} \in E$ and $e_{i,v}, e_{u,j} \notin E$, are randomly selected and switched to $e_{i,v}$ and $e_{u,j}$.

2.1.3.3 k-anonymity

As we discussed before, k-anonymity has been widely used to anonymize relational data. Similarly, many efforts have been spent to extend k-anonymity to graph data [38, 88, 156, 158, 160]. To defend against *neighborhood attacks*, Zhou and Pei proposed *k-Neighborhood Anonymity* (k-NA) for graph data [158]. k-NA is a two-step scheme. In the first step, the neighborhoods of all users (1-hop neighborhoods) are extracted and encoded in a concise way. In the second step, the users with similar neighborhoods are greedily grouped together until each group consists of at least k users, and then each group is anonymized such that any neighborhood has at least $k-1$ isomorphic neighborhoods in the same group. In another work, Liu and Terzi considered *degree attacks* and proposed *k-Degree Anonymity* (k-DA) for graph data, under which for each user, there exists at least $k-1$ other users with the degree [88]. k-DA also consists of two steps. First, based on the degree sequence of a graph, a new k-anonymous degree sequence (any degree appears at least k times in the sequence) is constructed. Second, an anonymized graph is constructed based on the k-anonymous degree sequence.

In [160], Zou et al. simultaneously considered four types of structural attacks to

[1]For simplicity and clarity, we use the same notation system as in existing work [152]-[159], [125]-[63], the structure of a graph dataset is modeled as a graph $G(V, E)$. More detailed information of this model will be provided later.

graph data: *neighborhood attacks* [158], *degree attacks* [88], *subgraph attacks* [27, 53], and *hub-fingerprint attacks* [53]. To defend against these attacks, they proposed *k-automorphism* (k-auto), under which for each user, there are always $k-1$ other symmetric users with respect to $k-1$ automorphic functions. To achieve k-auto, three techniques are developed, namely *graph partitioning*, *block alignment*, and *edge copy*. Another similar work is [38], where Cheng et al. proposed *k-isomorphism* (k-iso) to defend against structural attacks. Under k-iso, a graph is partitioned and anonymized into k disjoint subgraphs such that all these subgraphs are isomorphic. To ensure k-iso, both baseline and refined algorithms are designed. Furthermore, the authors demonstrated that k-iso is equivalent to k-auto in defending against user-deanonymization attacks.

In [156], Yuan et al. considered personalized privacy protection for anonymizing graph data in terms of both semantic and structural information. Based on the adversary's semantic and structural background knowledge, they customized three levels of privacy protection. Subsequently, different techniques are designed based on label generation (semantically) and noising edge/user addition (structurally) to achieve k-anonymity.

2.1.3.4 Aggregation/Class/Cluster based Anonymization

Another popular idea to protect graph data is to anonymize users into *clusters* (equivalently, *groups, classes*) [30, 53, 131]. In [53], Hay et al. proposed an *aggregation based graph anonymization* algorithm, which first partitions users and then describes the graph at the level of partitions. The anonymized graph consists of *supernodes*, each corresponding to the users in a partition, and *superedges*, indicating the edge densities among supernodes. Another work in the semantics level is [30], where Bhagat et al. designed an *interactive query-oriented* anonymization algorithm to partition a graph into classes with respect to users' attributes (labels). In [131], Thompson and

Yao first presented two clustering algorithms, named *bounded t-means* and *union-split* respectively, to classify users with similar rules into clusters. Subsequently, they proposed a *matching-based anonymization* scheme for graph data by strategically adding and removing edges according to users' inter-cluster connectivity.

2.1.3.5 Differential Privacy

Recently, there are some works that seek to enable differentially private graph data release. Aiming at protecting *edge/link privacy*, defined as *the privacy of users' relationship (e.g., friendship, contact, collaboration, email) in graph data*, in [122], Sala et al. introduced *Pygmalion*, a *differentially-private graph model*. In Pygmalioin, a graph is first modeled by dK-series, i.e., the degree distributions of connected components of some size K within a target graph. Subsequently, the dK-series is perturbed to meet ϵ-DP. Recently, to bypass many difficulties encountered when working with worst-case sensitivity [122], Proserpio presented a general platform, named *wPING*, for differentially private data analysis and publishing [117, 118]. Compared to previous solutions which scale up the magnitude of noise for challenging queries, wPING achieves better accuracy by scaling down the contributions of challenging records. Similar to [122], Wang and Wu also employed the dK-graph generation model for enforcing *edge DP* in graph anonymization. Another recent work for edge DP is [142], where Xiao et al. observed that, by estimating the connection probabilities among users instead of considering the edges directly, the noise scale enforced by edge DP can be significantly reduced. Following this observation, they proposed a *Hierarchical Random Graph* (HRG) model based scheme to meet edge DP.

2.1.3.6 Random Walk based Anonymization

In [97], Mittal et al. proposed a *Random Walk (RW) based anonymization* technique for preserving *link (edge) privacy*. By this technique, an edge between two users i and j is replaced by another edge between i and u, where u is the destination of a

random walk starting from j.

2.2 De-anonymization

In this section, we summarize state-of-the-art data de-anonymization attacks.

2.2.1 Relational Data De-anonymization

In [72], Lakshmanan et al. studied how safe anonymized data is with respect to protecting users' identities. They proposed various classes of belief functions to capture various degrees of partial information possessed by an adversary, and derived formulas for computing the expected number of cracks. In [103], Narayanan and Shmatikov presented a class of statistical de-anonymization attacks against high-dimensional micro-data. They further demonstrated the effectiveness of these attacks by successfully de-anonymizing the Netflix Prize dataset. In [42], Cormode studied the effectiveness of the *minimality attack*, which is an information inferring attack raised due to over-eager attempts to minimize the information lost by anonymization. Through careful analysis and experiments, they concluded that the impact of such attacks can be minimized.

In [101], Nanavati et al. presented an attack against reviewer anonymity. They showed that with access to a relatively small corpus of reviews, simple classification techniques from existing toolkits can successfully de-anonymize reviewers with reasonably high accuracy. In [41], Cormode studied the ability of an adversary to use data meeting privacy definitions to build an accurate classifier. They showed that private data can be accurately inferred even under DP. Furthermore, they observed that DP and ℓ-diversity are similar against classifier-based inference attack. In [95], Merener improved Narayanan and Shmatikov's work [103] on the de-anonymization of micro-data. They provided new results by considering cases where the auxiliary information has error and the dataset contains null values. Given auxiliary information of user's behavior, Unnikrishnan and Naini studied strategies for de-anonymizing

user statistics [134]. Particularly, they obtained an asymptotically optimal strategy when users' data following an independently and identically distribution model.

2.2.2 Graph Data De-anonymization

2.2.2.1 Seed-based De-anonymization

To de-anonymize graph data, it is intuitive to identify some users first as seeds. Then, the large scale de-anonymization is bootstrapped from these seeds. In [27], Backstrom et al. presented both active attacks and passive attacks to graph data. The active attacks are carried out in three steps. First, an adversary chooses a set of victims. Subsequently, the adversary creates some sybil accounts with edges linked to the victims, as well as a pattern of links among the sybil accounts before the data release. Finally, after data release, the adversary identifies the sybil accounts according to their structural pattern and then de-anonymizes the victims. In the passive attacks, an adversary is an internal user of the system and tries to de-anonymize the users around him after data release. The attacks in [27] have several limitations, e.g., they are not scalable, and they leverage sybil users that can be detected by modern sybil defense techniques [154, 155]. To improve the attacks in [27], Narayanan and Shmatikov presented a *scalable two-phase de-anonymization attack* to social networks [104]. In the first phase, some seed users are identified between the anonymized graph and the auxiliary graph. In the second phase, starting from the identified seeds, a self-reinforcing de-anonymization propagation process is iteratively conducted based on both graphs' structural characteristics, e.g., node degrees, nodes' eccentricity, edge directionality. Later, Narayanan employed a simplified version of the attack in [104] (using less de-anonymization heuristics) for link prediction [102]. Besides that, they also proposed a new simulated annealing-based weighted graph matching algorithm for the seed identifying phase (the first phase). In [108], Nilizadeh et al. further improved Narayanan and Shmatikov's attack by proposing a *community-enhanced* de-anonymization scheme of social networks. Specifically, the scheme first de-anonymizes

a social network at the community-level. Then, users within de-anonymized communities are further de-anonymized according to similar heuristics as in [104]. Actually, the community-level de-anonymization in [108] can also be applied to enhance other de-anonymization attacks [63, 64, 127, 151].

In [127], Srivatsa and Hicks presented three attacks to de-anonymize mobility traces, which can be modeled as contact graphs applying multiple preprocessing techniques (e.g., [115, 127]). Similar to Narayanan-Shmatikov attacks [102, 104], Srivatsa-Hicks attacks also consist of two phases, where the first phase is for seed identification and the second phase is for mapping (de-anonymization) propagation. To achieve mapping propagation, Srivatsa and Hicks proposed three heuristics based on *Distance Vector* (DV), *Randomized Spanning Trees* (RST), and *Recursive Subgraph Matching* (RSM). In [64], Ji et al. defined three similarity metrics, namely *structural similarity*, *relative distance similarity*, and *inheritance similarity*, and proposed two two-phase de-anonymization attack frameworks, named De-Anonymization (DA) and Adaptive De-Anonymization (ADA), which are workable when the auxiliary data only has partial overlap with the anonymized data.

In [69, 151], besides quantifying the de-anonymizability of graph data, the authors also proposed de-anonymization attacks. In [151], Yartseva and Grossglauser proposed a very simple *percolation-based de-anonymization algorithm* to graph data. Given a seed mapping set, the algorithm incrementally maps every pair of users (from the anonymized and auxiliary graphs respectively) with at least r neighboring mapped pairs, where r is a predefined mapping threshold. Another similar attack was presented by Korula and Lattanzi [69], which is also starting from a seed set and iteratively maps a pair of users with the most number of neighboring mapped pairs.

2.2.2.2 Seed-free De-anonymization

Recently, following another track, some powerful seed-free de-anonymization attacks on graph data are proposed. Using degrees and distances to other nodes as a nodes' fingerprints, Pedarsani et al. proposed a *Bayesian model based seed-free algorithm* for graph data de-anonymization [112]. Starting from nodes with the highest degree, the algorithm iteratively updates the fingerprints of all the nodes and performs a *maximum weighted bipartite graph matching* for de-anonymization. Another seed-free de-anonymization attack to graph data was presented by Ji et al. [63]. Unlike previous attacks, Ji et al.'s attack is an *optimization based single-phase cold start algorithm*. Following their theoretical analysis, their attack is iteratively conducted and self-reinforced with the objective of *minimizing the edge difference* between the anonymized graph and auxiliary graph.

2.2.2.3 Other Techniques

There are some other techniques that de-anonymize graph data, e.g., semantics based de-anonymization attacks [140], attacks to *ego graphs* [125], attacks against the link privacy of graph data [68]. By leveraging *web browser history stealing attack*, Wondracek et al. presented a de-anonymization attack to social networks based on users' *group membership information* [140]. Since we focus on structural de-anonymization attacks in this dissertation, we do not consider this kind of semantics based attacks. In [125], Sharad and Danezis studied the de-anonymization attacks to ego graphs with graph radius of one or two, which is a very special case of the general graph de-anonymization attacks studied in this dissertation. In [68], Korolova studied the link privacy leakage of anonymized social networks. In this dissertation, we focus on the de-anonymization attacks to the nodes (i.e., users) of graph data.

2.3 De-anonymizability Quantification

2.3.1 Seed-based Quantification

In [151], Yartseva and Grossglauser quantified the de-anonymizability of graph data by analyzing a percolation-based graph matching algorithm under the *Erdös-Rényi (ER) random graph model* $G(n, p)$ (a random graph consists of n nodes/users, and an edge exists between any pair of nodes with probability p). Under the ER model, the degree distribution of the considered graph data should follow the *Poisson distribution* [63, 107]. However, real world graph data may follow any distribution (e.g., many social networks follow the *power-law distribution*), and more importantly, seldom do we see any graph data following the Poisson distribution [63, 107]. Therefore, the quantification under the ER model is only mathematically meaningful but not practical. Nevertheless, it can shed light on more practical quantification. Another limitation of [151] is that it leverages seed-associated structural information for de-anonymizability quantification. In fact, as shown in [63, 113], graph data is de-anonymizable based solely on data's structural information, i.e., without seed.

Following the same direction, Korula and Lattanzi conducted another seed-based de-anonymizablity quantification of graph data under both the ER model and the *Preferential Attachment (PA) model* [69]. Again, several limitations make the quantification in [69] unpractical. First, as we mentioned before, the ER model is a theoretical model (i.e., it is not practical). Accordingly, the PA model is more practical compared to the ER. However, it still has some limitations, e.g., the existence of *self-loops*. Second, as in [151], the quantification in [69] only considers the structural information associated with seeds. Finally and more importantly, the quantification in [69] is valid under a strong assumption of existing dense seeds ($\Theta(\iota \cdot n)$ available seeds, $\iota \in (0, 1]$ is a constant), which is not true for real world de-anonymization attacks. Recently, Ji et al. quantified the seed-based de-anonymizability of social

networks [61] under both the ER model and a general *statistical graph model*. Compared to previous seed-based works, the quantification in [61] considers the structural information among anonymized users in addition to the structural information between anonymized users and seeds.

2.3.2 Seed-free Quantification

In [113], Pedarsani and Grossglauser quantified the de-anonymizability of graph data under the ER model. They showed that an anonymized graph is de-anonymizable when certain conditions on the structures of anonymized and auxiliary graphs are satisfied. Again, the quantification is under the mathematical ER model, which cannot be applied to real world graph data [63, 107]. Furthermore, for a de-anonymization attack, although it is improper to assume the availability of dense seeds, it is still reasonable to have some seed mappings as pre-knowledge [27, 64, 104, 127]. However, the quantification in [113] does not rely on seeds. Recently, Ji et al. improved the quantification in [113]. They quantified the *perfect* and *error-tolerated* de-anonymizability of graph data under a general *configuration model* [107], where the considered graph data can have an *arbitrary degree sequence*. Similar to [113], the quantification in [63] does not relay on seeds.

2.4 Research Evolution Summarization

As a fundamental and challenging problem space, data anonymization and de-anonymization have attracted a significant amount of attention from researchers. With the emergence of *big data*, this research becomes even more important and more challenging. Particularly, we summarize the evolution of data anonymization

Table 1: Anonymization techniques and de-anonymization attacks on relational (micro/tabular/set-valued) data. The *anonymization techniques that are italicized* are for set-valued data while the others are for micro/tabular data.

year	anonymization	de-anonymization
2001/2	k-anonymity [123, 128]	
2005	k-anonymity [24, 77]	Lakshmanan et al. [72]
2006	ℓ-diversity [90], (α, k)-anonymity [141], permutation [143], (X, Y)-privacy [136], k-AC [23], utility-aware [66, 76, 146], ϵ-DP [44], (ϵ, δ)-DP [45]	
2007	k-anonymity [111], t-closeness [83], (c, k)-safety [93], permutation [157], δ-presence [105], m-invariance [144]	
2008	m-invariance [145], utility-aware [32], *k^m-anonymity [130], (h, k, p)-coherence [147]*	Narayanan-Shmatikov [103]
2009	LKC-privacy [99], P-private [70] (ϵ, δ)-DP [94], *k-anonymity [57]*	
2010	predictive [34], ϵ-DP [54], *ρ-uncertainty [33]*	Cormode et al. [42]
2011	ϵ-DP [98]*[37]*	Nanavati et al. [101], Cormode [41]
2012	k-AC [91], ρ-DI [75], (ϵ, δ)-DP [84], *k-anonymity/ℓ-diversity [149], ϵ-DP [35, 36]*	Merener [95]
2013	b-matching [39], ϵ-DP [65], $(\mathbb{D}, \gamma)$-membership [85]	Unnikrishnan-Naini [134]
2014	ϵ-DP [82, 119]	

and de-anonymization research in Tables 1 and 2, from which we have the following observations.

- For anonymization techniques, following the seminal works of k-anonymity and DP, many schemes have been proposed to address the security and privacy concerns of both relational data and graph data in different scenarios, e.g., ℓ-diversity, (α, k)-anonymity, t-closeness, δ-presence, m-invariance, k^m-anonymity, (ϵ, δ)-DP, $(\mathbb{D}, \gamma)$-membership. This is mainly because these two privacy models provide formal methodologies for implementation, theoretical privacy guarantee, and moderate data utility preservation. Specifically, it seems that DP has attracted more research attention than k-anonymity recently. We conjecture that this is because DP is a relatively new technique and it provides an even stronger privacy guarantee than k-anonymity. However, proper application of DP for graph data anonymization is still in its infancy. Furthermore, the research of graph data anonymization started later than that of relational data, which is generally consistent with the evolution of computer data.

- With the popularity of graph data, more de-anonymization attacks on them have been presented as of recent. Similar to understanding the fundamental reasons that are responsible for the success of modern heuristic graph data de-anonymization attacks, researchers also began to conduct the research on quantifying the de-anonymizability of graph data.

- Most state-of-the-art graph data anonymization and de-anonymization schemes are based only on data's structural information. This is because (i) similar to the semantic information, the structure itself is also important information carried by graph data, which can be used for many data mining tasks and high level applications; (ii) many users in graph data have unique/quasi-unique topological structures, which can be used for identifying/quasi-identifying users; and

(*iii*) compared to semantic information, structure information is easier to obtain and analyze, which can be exploited for fast and effective de-anonymization attacks. Therefore, to protect graph data, researchers seek to anonymize the structural information, while to break the privacy of graph data, researchers try to exploit such information.

Table 2: Anonymization, de-anonymization, and quantification of graph data. **Bold techniques are anonymization algorithms or de-anonymization attacks based only data's structural information.**

year	anonymization	de-anonymization	quantification
2007		**Backstrom et al. [27]**	
2008	**Add/Del [152], Switch [152],** k**-NA [158],** k**-DA [88], aggregation [53]**	**Korolova et al. [68]**	
2009	k**-auto [160]**, class [30], **cluster [131]**	**Narayanan-Shmatikov [104]**	
2010	k**-iso [38],** k-anonymity [156]	Wondracek et al. [140]	
2011	ϵ**-DP [122]**	**Narayanan et al. [102]**	**Pedarsani-Grossglauser [113]**
2012	ϵ**-DP [117]**	**Srivatsa-Hicks [127]**	
2013	ϵ**-DP [137], randomization [97]**	**Pedarsani et al. [112], Sharad-Danezis [125]** **Yartseva-Grossglauser [151]**	**Yartseva-Grossglauser [151]**
2014	ϵ**-DP [118, 142]**	**Ji et al. [64], Korula-Lattanzi [69]** **Ji et al. [63], Nilizadeh et al. [108]**	**Korula-Lattanzi [69]** **Ji et al. [63]**
2015			**Ji et al. [61]**

CHAPTER III

SEED-FREE DE-ANONYMIZATION QUANTIFICATION

3.1 Introduction

Currently, to protect graph/structural data's privacy, the most common technique used is to anonymize data by removing the "*Personally Identifiable Information* (PII)" before releasing data. Unfortunately, this naive method is shown to be vulnerable to many De-Anonymization (DA) attacks [53, 84, 88]. Latterly, some sophisticated anonymization schemes to protect graph data privacy, e.g., k-anonymity and its variants [53, 84, 88], were designed[1]. They can protect the privacy of graph data to some extent. However, they are susceptible to emerging *Structure based De-Anonymization* (SDA) attacks[2] due to the limitations of the schemes (e.g., they are syntactic properties based) and the rich amount of information available to adversaries [27, 104, 127].

In SDA attacks, some auxiliary data (graphs) are employed to break the privacy of anonymized graph data based only on the structural information. The fact that the auxiliary data may come from either the same or a different domain/context with the anonymized data makes the attack powerful, e.g., using Flickr to de-anonymize Twitter [104], using Facebook to de-anonymize WiFi mobility traces [127]. Furthermore, the wide availability of auxiliary data makes the attack applicable and practical [104, 127].

The SDA attacks were initially presented in [27], where Backstrom et al. designed

[0]Without of specification, "de-anonymization" means "seed-free de-anonymization" in this chapter.

[1]Note that, the *differential privacy* [44] is well developed to protect the privacy of *interactive data release*. However, it cannot defense against *graph data DA attacks* which are designed to breach the privacy of *non-interactive data release* [44, 84, 104, 127].

[2]We use DA and SDA interchangeably.

both active and passive attacks to break the privacy of social network users. However, since the attacks in [27] leverage the success of a "sybil" attack before actual anonymized data publication, they are difficult to extend to large scale datasets. Later, Narayanan and Shmatikov designed a new SDA attack in [104], which successfully de-anonymizes a large scale directed social network by applying several heuristics such as eccentricity, edge directionality, reverse match, etc. In [127], Srivatsa and Hicks demonstrated that the privacy of three kinds of mobility traces can be compromised by SDA attacks. However, the attacks presented in [127] are only suitable for small datasets due to its computational infeasibility on finding a proper landmark mappings for large datasets. Note that each of the aforementioned attacks consist of two phases: a *landmark identification phase* and a *DA propagation phase*.

Although we already have some successful SDA practices [27, 104, 127], we do not have any *rigorous theoretical result under a general model* yet in answering why SDA attacks work. In [113], Pedarsani and Grossglauser quantified the privacy of anonymized graph data under the *Erdös-Rényi (ER) random graph model* $G(n, p)$ (every edge exits with identical probability p). However, this quantification is not suitable in practice since most, if not all, observed real world graph data (e.g., social networks, collaboration networks [16, 106, 107]) do not follow the ER model. Actually, they may follow the *power-law model*, *exponential model*, etc. [16, 106, 107]. Therefore, under a practical *general data model*, there are still some open problems in DA research: (*i*) *why can graph data be de-anonymized*? (*ii*) *what are the conditions for successful graph data DA*? and (*iii*) *what portion of users can be de-anonymized in a graph dataset*? To remedy the practice-theory gap, we study the *quantification*, *practice*, and *implications* of graph data DA in this chapter. Particularly, our contributions are as follows.

- To the best of our knowledge, this is the first work on quantifying graph data DA under a general data model. In our quantification, we answer several

fundamental open problems: why graph data can be de-anonymized based only on the topological information (the inherent reason for the success of existing SDA practices)? what are the conditions for *perfect* and $(1-\epsilon)$-*perfect DA*, where ϵ is the *error* tolerated by a DA scheme? what portion of users can be de-anonymized in a graph dataset? Thus, we close the gap between graph data DA practice and theory.

- We conduct the first large-scale study on the de-anonymizability of 26 real world graph datasets, including social networks, location based mobility traces and social networks, collaboration networks, communication networks (Email, WikiTalk), autonomous system graph data, peer-to-peer network data, etc. Based on our study, we find *all* the considered graph datasets are de-anonymizable perfectly or partially. We also quantitatively show the conditions for perfect and $(1-\epsilon)$-perfect DA and what portion of users can be de-anonymized for the 26 datasets.

- Following our quantification, we present a novel *Optimization based DA* (ODA) attack. Different from existing SDA attacks [27, 104, 127], ODA is a *single-phase cold start* algorithm without any requirement on priori knowledge, e.g., landmark mappings. We also examine ODA on real datasets Gowalla (196,591 users and 950,327 edges) and Google+ (4,692,671 users and 90,751,480 edges). The results demonstrate that about $77.7\% - 83.3\%$ of the users in Gowalla and $86.9\% - 95.5\%$ of the users in Google+ are de-anonymizable, which implies SDA is implementable and powerful in practice.

- Finally, we discuss some implications of this work according to our graph DA quantification and the ODA attack. We further provide some general suggestions for future *secure data publishing*.

The rest of this chapter is organized as follows. In Section 3.2, we give the data

and attack models. In Section 3.3, we theoretically quantify perfect and $(1-\epsilon)$-perfect DA attacks under a general data model, followed by a large-scale evaluation on 26 diverse real world graph datasets in Section 3.4. In Section 3.5, we present a novel optimization based DA attack with theoretical and experimental analysis. We discuss the implications of our DA quantification and ODA attack in Section 3.6. The chapter is concluded and future work is addressed in Section 3.7.

3.2 System Model

In this chapter, we focus on quantifying and analyzing the DA attack (vulnerability) on anonymized graph data, which could be social data released by social network operators, e.g., Google+ [49], Facebook [16], Twitter [16], and/or mobility data generated by mobile devices, e.g., WiFi and Bluetooth traces [127], instant message contacts [127], email networks [16], classical longitude-latitude spatiotemporal traces [16, 115]. In the following subsection, we formally define the anonymized and auxiliary data models, as well as the attack model.

3.2.1 Data Model

It is straightforward to model social data using graphs, where nodes represent users and edges/links indicate the social relationships (*friendship*, *contact*, *following*) among users. For the mobility data generated by users (users' devices), they can also be modeled by contact graphs according to recently proposed techniques [115, 127]. Furthermore, it has been shown that a contact graph derived from mobility data has strong correlation (similarity) with the social graph of the same group of users that generated them [115, 127]. Therefore, we model the anonymized graph data by a graph $G^a = (V^a, E^a)$, where $V^a = \{i | i \text{ is an } \textit{anonymized} \text{ user}\}$ is the user set and $E^a = \{e^a_{i,j} | \text{ there is a relationship (friend, contact, etc.) between } i \in V^a \text{ and } j \in V^a\}$ is the edge/relationship set. In reality, it is possible that a graph dataset corresponds

to a directed graph, e.g., Twitter. However, for simplicity and without loss of generality, we assume G^a as an undirected graph. Note that, the designed algorithm in this chapter can be extended to the directed scenario directly. For $i \in V^a$, its neighborhood is defined as $N_i^a = \{j | \exists e_{i,j}^a \in E^a\}$ and we denote the cardinality of N_i^a as $|N_i^a|$, i.e., the degree of i.

The auxiliary data is also assumed to be graph data, e.g., a social network compromising users overlapped with that in the anonymized graph data [104, 127]. Furthermore, the auxiliary data is easily obtainable by multiple means such as academic and government data mining, advertising, third-party applications, data aggregation, online crawling, etc. Successful examples can be found in [104, 115, 127, 140]. Consequently, the auxiliary data is also modeled by a graph $G^u = (V^u, E^u)$, where $V^u = \{i$ is a *known* user$\}$ and $E^u = \{e_{i,j}^u |$ there is a relationship (friend, contact, etc.) between $i \in V^u$ and $j \in V^u\}$. Similarly, the neighborhood of $i \in V^u$ is defined as $N_i^u = \{j | \exists e_{i,j}^u \in E^u\}$.

3.2.2 De-anonymization Attack

Given G^a and G^u, a DA attack can be formally defined as a *mapping*:

$$\sigma : V^a \to V^u. \tag{1}$$

For $\forall i \in V^a$, its mapping under σ is $\sigma(i) \in V^u \cup \{\perp\}$, where $\perp$ is a special *not existing indicator*. Similarly, for $\forall e_{i,j}^a \in E^a$, $\sigma(e_{i,j}^a) = e_{\sigma(i),\sigma(j)}^u \in E^u \cup \{\perp\}$. Under σ, a successful DA on $i \in V^a$ is defined as

$$\sigma(i) = \begin{cases} i', & \text{if } i' \in V^u \text{ and } i \text{ and } i' \text{ correspond to the same user;} \\ \perp, & \text{otherwise.} \end{cases} . \tag{2}$$

For other cases, the DA on i fails. Consequently, the objective of a DA attack is to successfully de-anonymize as many users in V^a as possible.

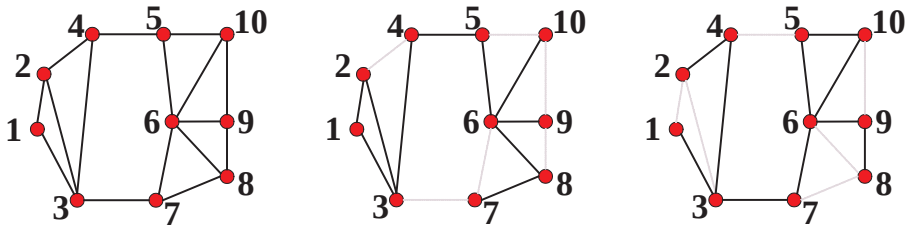
3.3 De-anonymization Quantification

In this section, given G^a and G^u, we quantify a DA attack under an *arbitrary graph distribution* in multiple scenarios. Particularly, we study the condition on the structure of anonymized data under which a successful DA attack can be conducted. Note that, our quantification is aiming at providing a theoretical foundation on understanding the success of recent heuristic SDA practices [104, 127]. We theoretically demonstrate that even without any further (e.g., semantic) knowledge, perfect or $(1-\epsilon)$-perfect DA attacks can be implemented when some structural conditions on the underlaying graph corresponding to G^a and G^u are satisfied.

3.3.1 Preliminaries

To make the quantification and proof tractable and convenient, we make some assumptions and definitions. First, we assume $V^a = V^u$, i.e., the auxiliary data and the anonymized data are corresponding to the same group of users [104, 113, 127]. This does not mean that we know any priori correct mapping from V^a to V^u. Furthermore, this assumption is reasonable since one cannot be expected to use G^u to de-anonymize G^a if they correspond to different groups of users. It is possible that the auxiliary data only has some overlap with the anonymized data instead of corresponding to the exactly same group of users. This fact does not limit our theoretical analysis since we can either (i) apply the quantification to the overlap part, or (ii) redefine $V^a_{new} = V^a \cup (V^u \setminus V^a)$ and $V^u_{new} = V^u \cup (V^a \setminus V^u)$, i.e. adding the non-overlapped users to V^a and V^u respectively as isolated users (with degree 0), and apply the analysis to $G^a = (V^a_{new}, E^a)$ and $G^u = (V^u_{new}, E^u)$. Without of causing any confusion, we assume $V^a = V^u$ in the rest of this section.

Second, similar to the methodology in [113], for the users in V^a (or, V^u), we assume that there exists a conceptual underlying graph $G = (V, E)$ with $V = V^a = V^u$ and E consisting of the true relationships among users in V. Consequently, G^a



(a) G (b) G^a (c) G^u

Figure 3: Edge/relationship projection. Only black edges appear in G^a/G^u.

and G^u can be viewed as the physically observable *projections* of G on particular relationships, e.g., "friendship" relationship on Facebook, "circle" relationship on Google+, "follow" relationship on Twitter, "co-occurrence" relationship in Gowalla, "coauthor" relationship in DBLP. The projection from G to G^a is characterized by an *edge/relationship projection process* [113]: (i) $V^a = V$; and (ii) $\forall e_{i,j} \in E$, $e_{i,j}$ is appeared in E^a with probability p_a, i.e., $\Pr(e_{i,j} \in E^a | e_{i,j} \in E) = p_a$. Similarly, the projection from G to G^u can be characterized by another *edge/relationship projection process* with probability p_u. For instance, we show a projection from G to G^a/G^u in Fig. 3. Furthermore, we assume both projection processes are *independent and identically distributed* (*i.i.d.*). Note that, (i) although the assumption on the existence of a conceptual underlying graph and the projection process makes the quantification problem theoretically tractable, it is still a challenging issue in practice; and (ii) assuming G^a and G^u are projected from an underlying network implies G^a and G^u have a strong structural correlation. Intuitively, this assumption is reasonable since they correspond to the same group of users and the empirical results in [104, 127] also supports such strong structural correlation.

Evidently, based on the above assumptions, we have $n!$ possible DA schemes $\sigma : V^a \to V^u$ to de-anonymize G^a, among which the only one *perfect DA scheme* ($\forall i \in V^a$, i is successfully de-anonymized) is denoted by σ_0.

3.3.2 Model and Formalization

Now, given G, we denote $|V| = n$ and $|E| = m$. Let $V = \{1, 2, \cdots, n\}$ and d_i be the degree of $i \in V$. Then, we define $\mathbf{D} =< d_1, d_2, \cdots, d_n >$ as the degree sequence of the nodes (users) in V. Furthermore, let Δ_1 and Δ_2 (resp., δ_1 and δ_2) be the *maximum* and *second maximum* (resp., *minimum* and *second minimum*) degrees of G, respectively. In [113], Pedarsani and Grossglauser quantified the privacy of G when G is an ER random graph $G(n, p)$ [3]. The $G(n, p)$ model is very useful as a source of insight into the study of graph data, e.g., social networks [107, 113]. However, the degree distribution of $G(n, p)$ tends to follow the Poisson distribution, which is quite different from the degree distributions of most, if not all, observed real world graph data (e.g., social networks, collaboration networks, mobility based contact networks.) [106, 107]. Actually, the degree distribution of real world graph data (represented by graphs) may follow any distribution such as the power-law distribution, exponential distribution, etc. [106, 107]. Therefore, it is significant to understand and quantify a DA attack (or the privacy and vulnerability) for graph data under an *arbitrary degree distribution*. To this end, we characterize G by a generalized graph model, the *configuration model* [107]. Under the configuration model, a graph is specified by an arbitrary degree sequence $\mathbf{D}$ rather than a particular degree distribution. Since $\mathbf{D}$ is an arbitrary degree sequence, $\mathbf{D}$ can follow an arbitrary degree distribution observed in real world data [107].

Let $p_{i,j}$ be the probability of existing an edge between $i, j \in V$. Then, we have

$$p_{i,j} = \frac{d_i d_j}{2m - 1} \underset{m \to \infty}{\overset{\text{as}}{\simeq}} \frac{d_i d_j}{2m}, \tag{3}$$

which is a key property of the configuration model [107]. From $p_{i,j}$, it is more likely of an existing edge (relationship) between two users with high degrees. Based on $p_{i,j}$, we define $l = \min\{p_{i,j} | i, j \in V, i \neq j\}$ and $h = \max\{p_{i,j} | i, j \in V, i \neq j\}$, i.e., l and

[3] Based on the projection process, G^a and G^u are also ER random graphs $G(n, p \cdot p_a)$ and $G(n, p \cdot p_u)$, respectively.

h are the lower and upper bounds of $p_{i,j}$ respectively. Then, given G with arbitrary degree distribution, we have $l \geq \frac{\delta_1 \delta_2}{2m-1}$ and $h \leq \frac{\Delta_1 \Delta_2}{2m-1}$.

Finally, given any DA scheme $\sigma = \{(i, i') | 1 \leq i, i' \leq n, i \in V^a, i' \in V^u\} \subseteq V^a \times V^u$, we define the *De-anonymization Error* (DE) on a user mapping $(i, i') \in \sigma$ as

$$\psi_{i,i'} = |N_i^a \setminus N_{i'}^u| + |N_{i'}^u \setminus N_i^a|, \tag{4}$$

which measures the neighborhoods' difference between i in G^a and i' in G^u under the particular σ. Then, we define the overall DE for a particular σ as

$$\Psi_\sigma = \sum_{(i,i') \in \sigma} \psi_{i,i'}. \tag{5}$$

Taking G^a and G^u shown in Fig. 3 as an example, the DE of the perfect DA scheme σ_0 is $\Psi_{\sigma_0} = 20$. For another DA scheme $\sigma = (\sigma_0 \setminus \{(4,4),(5,5)\}) \cup \{(4,5),(5,4)\}$ (users 4 and 5 are incorrectly de-anonymized to each other), its DE is $\Psi_\sigma = 28$. In the following subsections, *we quantify a DA attack by studying the conditions on G and the projection process under which perfect and $(1-\epsilon)$-perfect DA attacks can be conducted*. Equivalently, *we study the conditions on G and the projection process such that the perfect/$(1-\epsilon)$-perfect DA scheme minimizes DE* (*mathematically, this implies a perfect/$(1-\epsilon)$-perfect DA scheme can be obtained since the number of DA schemes is bounded*).

3.3.3 Perfect De-anonymization

Now, we quantify the conditions for perfect DA attacks. Then, we extend to the scenario of quantifying $(1-\epsilon)$-perfect DA attacks. Some useful properties of the *binomial distribution* that will be used in the proofs are as follows.

Lemma 1. *(i) Let $X \sim \mathbf{B}(n_1, p)$ and $Y \sim \mathbf{B}(n_2, p)$ be independent binomial variables. Then, $X+Y$ is again a binomial variable and $X+Y \sim \mathbf{B}(n_1+n_2, p)$; (ii) [113] Let X and Y be two binomial random variables with means λ_x and λ_y, respectively. Then,*

when $\lambda_x > \lambda_y$,

$$\Pr(X - Y \leq 0) \leq 2\exp(-\frac{(\lambda_x - \lambda_y)^2}{8(\lambda_x + \lambda_y)}). \tag{6}$$

3.3.3.1 Same Projection Probability

First, we consider the scenario that the projection processes from G to G^a and G^u are characterized by the same probability $\wp$, i.e., $p_a = p_u = \wp$. Let

$$f_\wp = \frac{\wp[l(1 - h\wp) - h(1 - \wp)]^2}{2(l(1 - h\wp) + h(1 - \wp))} \tag{7}$$

be a variable depending on $\wp$. Then, we have the following Theorem 1 which indicates the conditions on $\wp$ and $f_\wp$ such that it is *asymptotically almost surely* (*a.a.s.*)[4] that $\Psi_\sigma \geq \Psi_{\sigma_0}$ for any DA scheme $\sigma \neq \sigma_0$.

Theorem 1. *For any* $\sigma \neq \sigma_0$*, let* k *be the number of different mappings between* σ *and* σ_0*, i.e., the number of incorrect mappings in* σ*. Then,* $2 \leq k \leq n$ *and* $\Pr(\Psi_\sigma \geq \Psi_{\sigma_0}) \underset{n\to\infty}{\to} 1$ *when* $\wp > \frac{h-l}{h-hl}$ *and* $f_\wp = \Omega(\frac{2\ln n+1}{kn})$.

Proof: Since k is the number of incorrect mappings in $\sigma \neq \sigma_0$, $2 \leq k \leq n$ is evidently. For convenience of proof, let σ_k be a DA scheme that has k incorrect (unsuccessful) mappings. Under σ_k, let $V_k \subseteq V$ be the set of incorrectly de-anonymized users[5], $\mathcal{E}_k = \{e_{i,j} | i \in V_k \text{ or } j \in V_k\}$ be the set of all possible edges adjacent to at least one user in V_k, $\mathcal{E}_\tau = \{e_{i,j} | i, j \in V_k, (i,j) \in \sigma_k, \text{ and } (j,i) \in \sigma_k\}$ be the set of all possible edges corresponding to *transposition mappings*[6] in σ_k, and $\mathcal{E} = \{e_{i,j} | 1 \leq i \neq j \leq n\}$ be the set of all possible edges on V. Furthermore, define $m_k = |\mathcal{E}_k|$ and $m_\tau = |\mathcal{E}_\tau|$. Then, we have $|V_k| = k$, $m_k = \binom{k}{2} + k(n-k)$, $m_\tau \leq \frac{k}{2}$ since there are at most $\frac{k}{2}$ transposition mappings in σ_k, $|\mathcal{E}| = \binom{n}{2}$, and $\forall e_{i,j} \in \mathcal{E}$, $\Pr(e_{i,j} \in E) = p_{i,j} = \frac{d_i d_j}{2m-1}$.

[4] *Asymptotically almost surely* (*a.a.s.*) implies that as $n \to \infty$, with probability goes to 1 an event happens.

[5] Without of causing any confusion, we use V, V^a, and V^u interchangeably since $V = V^a = V^u$.

[6] If both mappings (i,j) and (j,i) are in σ_k, then $\{(i,j),(j,i)\}$ is called a *transposition mapping*, i.e., two users are incorrectly de-anonymized to each other.

Now, we quantify Ψ_{σ_0} stochastically. During the quantification, we employ a widely used technique in graph theory [31, 113]. That is, to quantify Ψ_{σ_0}, we considering the DE caused by the projection of each edge rather than considering the mapping directly. $\forall e_{i,j} \in \mathcal{E}$, if it appears in E and is projected to either G^a or G^u but not both during the edge projection process, then according to the definition of DE, it will cause a DE of 2. Consequently, the DE caused by $e_{i,j}$ satisfies a binomial distribution $\mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp))$. Furthermore, since the projection process is *i.i.d.* and considering Lemma 1, we have

$$\Psi_{\sigma_0} = \sum_{(t,t') \in \sigma_0} \psi_{t,t'} \tag{8}$$

$$\sim \sum_{e_{i,j} \in \mathcal{E}} \mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp)) \tag{9}$$

$$= \mathbf{B}(\sum_{e_{i,j} \in \mathcal{E}} 2, 2p_{i,j} \cdot \wp(1-\wp)). \tag{10}$$

When quantify Ψ_{σ_k}, we consider three cases respectively.

• *Case 1*: for $\forall e_{i,j} \in \mathcal{E} \setminus \mathcal{E}_k$, the DE caused by $e_{i,j}$ during the projection process also satisfies the binomial distribution $\mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp))$ since $i, j \in V \setminus V_k$ (i.e., i, j are successfully de-anonymized under σ_k).

• *Case 2:* for $\forall e_{i,j} \in \mathcal{E}_k \setminus \mathcal{E}_\tau$, it will be mapped to some other possible edge $\sigma_k(e_{i,j}) = e_{\sigma_k(i),\sigma_k(j)} \in \mathcal{E}$ since $e_{i,j} \notin \mathcal{E}_\tau$ and at least one of i and j is incorrectly de-anonymized under σ_k. Therefore, in this case, the DE caused by $e_{i,j}$ during the projection process satisfies binomial distribution $\mathbf{B}(2, p_{i,j} \cdot \wp(1 - p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)} \cdot \wp(1 - p_{i,j}\wp))$.

• *Case 3:* for $\forall e_{i,j} \in \mathcal{E}_\tau$, since it corresponds to a transposition mapping, the DE caused by $e_{i,j}$ during the projection process also satisfies the binomial distribution $\mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp))$.

In summary, we have

$$\Psi_{\sigma_k} = \sum_{(t,t')\in\sigma_k} \psi_{t,t'} \tag{11}$$

$$\sim \sum_{e_{i,j}\in\mathcal{E}\setminus\mathcal{E}_k} \mathbf{B}(2, 2p_{i,j}\cdot\wp(1-\wp)) \tag{12}$$

$$+ \sum_{e_{i,j}\in\mathcal{E}_k\setminus\mathcal{E}_\tau} \mathbf{B}(2, p_{i,j}\cdot\wp(1-p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)}\cdot\wp(1-p_{i,j}\wp)) \tag{13}$$

$$+ \sum_{e_{i,j}\in\mathcal{E}_\tau} \mathbf{B}(2, 2p_{i,j}\cdot\wp(1-\wp)) \tag{14}$$

$$\underset{stochastically}{\geq} \sum_{e_{i,j}\in\mathcal{E}\setminus\mathcal{E}_k} \mathbf{B}(2, 2p_{i,j}\cdot\wp(1-\wp)) \tag{15}$$

$$+ \sum_{e_{i,j}\in\mathcal{E}_k\setminus\mathcal{E}_\tau} \mathbf{B}(2, p_{i,j}\cdot\wp(1-p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)}\cdot\wp(1-p_{i,j}\wp)) \tag{16}$$

$$= \mathbf{B}(\sum_{e_{i,j}\in\mathcal{E}\setminus\mathcal{E}_k} 2, 2p_{i,j}\cdot\wp(1-\wp)) \tag{17}$$

$$+ \mathbf{B}(\sum_{e_{i,j}\in\mathcal{E}_k\setminus\mathcal{E}_\tau} 2, p_{i,j}\cdot\wp(1-p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)}\cdot\wp(1-p_{i,j}\wp)). \tag{18}$$

Now, define $X \sim \mathbf{B}(\sum_{e_{i,j}\in\mathcal{E}_k\setminus\mathcal{E}_\tau} 2, p_{i,j}\cdot\wp(1-p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)}\cdot\wp(1-p_{i,j}\wp))$ and $Y \sim \mathbf{B}(\sum_{e_{i,j}\in\mathcal{E}_k} 2, 2p_{i,j}\cdot\wp(1-\wp))$. Let λ_x and λ_y by the mean values of X and Y, respectively. Thus, we have

$$\lambda_x = (\sum_{e_{i,j}\in\mathcal{E}_k\setminus\mathcal{E}_\tau} 2)\cdot[p_{i,j}\cdot\wp(1-p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)}\cdot\wp(1-p_{i,j}\wp)] \tag{19}$$

$$= 2(m_k - m_\tau)[p_{i,j}\cdot\wp(1-p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)}\cdot\wp(1-p_{i,j}\wp)] \tag{20}$$

$$\geq 2(m_k - m_\tau)\cdot[2l\wp(1-h\wp)] \tag{21}$$

$$= 4l\wp(1-h\wp)(m_k - m_\tau) \tag{22}$$

and

$$\lambda_y = (\sum_{e_{i,j}\in\mathcal{E}_k} 2)\cdot[2p_{i,j}\cdot\wp(1-\wp)] \tag{23}$$

$$\leq 4h\wp(1-\wp)m_k. \tag{24}$$

Then, $\forall \sigma_k$ $(k \in [2, n])$,

$$\Pr(\Psi_{\sigma_k} - \Psi_{\sigma_0} \leq 0) \underset{stochastically}{\leq} \Pr(\sum_{e_{i,j} \in \mathcal{E} \setminus \mathcal{E}_k} \mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp)) \tag{25}$$

$$+ \sum_{e_{i,j} \in \mathcal{E}_k \setminus \mathcal{E}_\tau} \mathbf{B}(2, p_{i,j} \cdot \wp(1 - p_{\sigma_k(i),\sigma_k(j)}\wp) + p_{\sigma_k(i),\sigma_k(j)} \cdot \wp(1 - p_{i,j}\wp)) \tag{26}$$

$$- \sum_{e_{i,j} \in \mathcal{E}} \mathbf{B}(2, 2p_{i,j} \cdot \wp(1-\wp))) \tag{27}$$

$$\underset{stochastically}{=} \Pr(X - Y \leq 0) \tag{28}$$

We now derive the upper bound on $\Pr(X - Y \leq 0)$. Since $\wp > \frac{h-l}{h-hl}$, $m_\tau \leq \frac{k}{2}$, and $m_k = \binom{k}{2} + k(n-k)$,

$$\wp > \frac{h-l}{h-hl} = \frac{(h-l)m_k}{(h-hl)m_k} \underset{n \to \infty}{\simeq} \frac{(h-l)m_k + lm_\tau}{(h-hl)m_k + lhm_\tau} \tag{29}$$

$$\Leftrightarrow \wp[(h-hl)m_k + lhm_\tau] > (h-l)m_k + lm_\tau \tag{30}$$

$$\Leftrightarrow l(m_k - m_\tau) - lh(m_k - m_\tau)\wp > hm_k - hm_k\wp \tag{31}$$

$$\Leftrightarrow 4l\wp(1-h\wp)(m_k - m_\tau) > 4h\wp(1-\wp)m_k \tag{32}$$

$$\Rightarrow \lambda_x > \lambda_y. \tag{33}$$

Applying Lemma 1 and considering that $f_{\wp} = \Omega(\frac{2\ln n+1}{kn})$, we have

$$\Pr(X - Y \leq 0) \leq 2\exp(-\frac{(\lambda_x - \lambda_y)^2}{8(\lambda_x + \lambda_y)}) \tag{34}$$

$$\leq 2\exp(-\frac{(4l\wp(1-h\wp)(m_k - m_\tau) - 4h\wp(1-\wp)m_k)^2}{8(4l\wp(1-h\wp)(m_k - m_\tau) + 4h\wp(1-\wp)m_k)}) \tag{35}$$

$$= 2\exp(-\frac{\wp(l(1-h\wp)(m_k - m_\tau) - h(1-\wp)m_k)^2}{2(l(1-h\wp)(m_k - m_\tau) + h(1-\wp)m_k)}) \tag{36}$$

$$= 2\exp(-\frac{\wp[(l(1-h\wp) - h(1-\wp))m_k - l(1-h\wp)m_\tau]^2}{2((l(1-h\wp) + h(1-\wp))m_k - l(1-h\wp)m_\tau)}) \tag{37}$$

$$\underset{n \to \infty}{\simeq} 2\exp(-\frac{\wp[l(1-h\wp) - h(1-\wp)]^2 m_k}{2(l(1-h\wp) + h(1-\wp))}) \tag{38}$$

$$= 2\exp(-f_{\wp} m_k) \tag{39}$$

$$= 2\exp(-\Omega(\frac{2\ln n + 1}{kn}) \cdot (\binom{k}{2} + k(n-k))) \tag{40}$$

$$\leq 2\exp(-2\ln n - 1) \tag{41}$$

$$\leq \frac{1}{n^2}. \tag{42}$$

Define $\zeta(2) = \sum_{n>0} \frac{1}{n^2}$. Then, $\zeta(2)$ is the *Euler-Riemann zeta function* with parameter 2 and thus $\zeta(2) = \frac{\pi^2}{6} < \infty$. Consequently, according to the *Borel-Cantelli Lemma*, it is *a.a.s.* that $X \geq Y$. It follows that it is *a.a.s.* that $\Psi_{\sigma_k} \geq \Psi_{\sigma_0}$ for $2 \leq k \leq n$, i.e., $\Pr(\Psi_\sigma \geq \Psi_{\sigma_0}) \to 1$ for any $\sigma \neq \sigma_0$. □

In Theorem 1, we quantified the condition on $\wp, l$, and h under which the perfect DA scheme σ_0 will cause less DE than any other given DA scheme $\sigma \neq \sigma_0$. To guarantee the *uniqueness* of σ_0 (i.e., σ_0 is *the one and the only one* DA scheme introducing the least DE), intuitively, stronger conditions on $\wp, l$, and h are required. We quantify such conditions in Theorem 2.

Theorem 2. *Let* **E** *be the event that* there exists at least one DA scheme $\sigma \neq \sigma_0$ such that $\Psi_\sigma \leq \Psi_{\sigma_0}$. *When* $\wp > \frac{h-l}{h-hl}$ *and* $f_{\wp} = \Omega(\frac{(k+3)\ln n+1}{kn})$, *where* $2 \leq k \leq n$, $\Pr(\mathbf{E}) \to 0$, *i.e., it is* a.a.s. *that there exists no DA scheme* σ *such that* $\sigma \neq \sigma_0$ *and* $\Psi_\sigma \leq \Psi_{\sigma_0}$.

Proof: Let $\mathbf{E}_\sigma$ be the event that $\Psi_\sigma \leq \Psi_{\sigma_0}$. Then,

$$\Pr(\mathbf{E}) = \Pr(\bigcup_{\sigma} \mathbf{E}_\sigma) = \Pr(\bigcup_{k=2}^{n} \bigcup_{\sigma_k} \mathbf{E}_{\sigma_k}). \tag{43}$$

Let ϱ_k be the number of DA schemes having k incorrect mappings. Then, $\varrho_k = \binom{n}{k} \cdot !k \leq n^k$, where $!k$ is the subfactorial of k [113, 120]. Then, considering that $f_\wp = \Omega(\frac{(k+3)\ln n+1}{kn})$ and based on *Boole's inequality* and the proof of Theorem 1, we have

$$\begin{aligned}
\Pr(\mathbf{E}) &= \Pr(\bigcup_{k=2}^{n} \bigcup_{\sigma_k} \mathbf{E}_{\sigma_k}) && (44)\\
&\leq \sum_{k=2}^{n} \varrho_k \cdot \Pr(\Psi_{\sigma_k} \leq \Psi_{\sigma_0}) && (45)\\
&\leq \sum_{k=2}^{n} n^k \cdot 2\exp(-f_\wp m_k) && (46)\\
&= \sum_{k=2}^{n} 2\exp(k\ln n - f_\wp m_k) && (47)\\
&= \sum_{k=2}^{n} 2\exp(k\ln n - \Omega(\frac{(k+3)\ln n+1}{kn})(\binom{k}{2} + k(n-k))) && (48)\\
&\underset{n\to\infty}{\leq} \sum_{k=2}^{n} 2\exp(k\ln n - (k+3)\ln n - 1) && (49)\\
&\leq \sum_{k=2}^{n} \frac{1}{n^3} && (50)\\
&\leq \frac{1}{n^2}. && (51)
\end{aligned}$$

Again, since $\zeta(2) = \sum_{n>0} = \frac{\pi^2}{6} < \infty$, it is *a.a.s.* that $\Pr(\mathbf{E}) \to 0$ based on the *Borel-Cantelli Lemma*, i.e., it is *a.a.s.* that there exists no DA scheme such that $\sigma \neq \sigma_0$ and $\Psi_\sigma \leq \Psi_{\sigma_0}$. □

From Theorem 2, although we seek a stronger result, the condition on $\wp$ is the same as in Theorem 1 and the condition on $f_\wp$ only has an increase of order $\Theta(k)$. Based on Theorem 2, if $\wp > \frac{h-l}{h-hl}$ and $f_\wp = \Omega(\frac{(k+3)\ln n+1}{kn})$, the perfect DA scheme causes the least DE. Furthermore, the number of possible DA schemes is upper-bounded.

Therefore, when the conditions on $\wp$ and $f_\wp$ are satisfied, G^a can mathematically be perfectly de-anonymized by G^u based on the structure information only.

3.3.3.2 Different Projection Probabilities

In this subsection, we quantify the conditions on p_a, p_u, l, and h when $p_a \neq p_u$ for structure based perfect DA attacks. Let

$$g_{p_a,p_u} = \frac{p_a p_u}{p_a + p_u} \tag{52}$$

and

$$f_{p_a,p_u} = \frac{(l(p_a + p_u - 2hp_a p_u) - h(p_a + p_u - 2p_a p_u))^2}{4(l(p_a + p_u - 2hp_a p_u) + h(p_a + p_u - 2p_a p_u))} \tag{53}$$

be two variables depending on p_a and p_u. Then, we have the following theorem quantifying the conditions on $g_{p_a,p_u}, f_{p_a,p_u}, l$, and h under which it is *a.a.s.* $\Psi_\sigma \geq \Psi_{\sigma_0}$ for any $\sigma \neq \sigma_0$. Note that, without causing any confusion, we consistently employ the same notations as in Theorems 1 and 2 in the following of this section.

Theorem 3. *When* $g_{p_a,p_u} > \frac{h-l}{2(h-lh)}$ *and* $f_{p_a,p_u} = \Omega(\frac{2\ln n+1}{kn})$, $\Pr(\Psi_\sigma \geq \Psi_{\sigma_0}) \to 1$ *for any* $\sigma \neq \sigma_0$.

Proof Sketch: Basically, this theorem can be proven following a similar idea as in Theorem 1. The main differences lies in the probabilities during the edge/relationship projection process. Now,

$$\Psi_{\sigma_0} \sim \mathbf{B}(\sum_{e_{i,j}\in\mathcal{E}} 2, p_{i,j}(p_a(1-p_u) + p_u(1-p_a))) \tag{54}$$

and

$$\Psi_{\sigma_k} \underset{stochastically}{\geq} \mathbf{B}(\sum_{e_{i,j}\in\mathcal{E}\setminus\mathcal{E}_k} 2, p_{i,j}(p_a(1-p_u) + p_u(1-p_a))) \tag{55}$$

$$+ \mathbf{B}(\sum_{e_{i,j}\in\mathcal{E}_k\setminus\mathcal{E}_\tau} 2, p_{i,j}\cdot p_a(1-p_{\sigma_k(i),\sigma_k(j)}p_u) + p_{\sigma_k(i),\sigma_k(j)}\cdot p_u(1-p_{i,j}p_a)). \tag{56}$$

Define $X \sim \mathbf{B}(\sum_{e_{i,j}\in\mathcal{E}_k\setminus\mathcal{E}_\tau} 2, p_{i,j}\cdot p_a(1-p_{\sigma_k(i),\sigma_k(j)}p_u)+p_{\sigma_k(i),\sigma_k(j)}\cdot p_u(1-p_{i,j}p_a))$ and $Y \sim \mathbf{B}(\sum_{e_{i,j}\in\mathcal{E}_k} 2, p_{i,j}(p_a(1-p_u)+p_u(1-p_a)))$. Then, we have

$$\lambda_x = 2(m_k - m_\tau)\cdot(p_{i,j}\cdot p_a(1-p_{\sigma_k(i),\sigma_k(j)}p_u)+p_{\sigma_k(i),\sigma_k(j)}\cdot p_u(1-p_{i,j}p_a)) \tag{57}$$

$$\geq 2l(p_a+p_u-2hp_ap_u)(m_k-m_\tau). \tag{58}$$

and

$$\lambda_y = 2m_k\cdot p_{i,j}(p_a(1-p_u)+p_u(1-p_a)) \tag{59}$$

$$\leq 2h(p_a+p_u-2p_ap_u)m_k. \tag{60}$$

Since $g_{p_a,p_u} > \frac{h-l}{2(h-lh)}$,

$$g_{p_a,p_u} = \frac{p_ap_u}{p_a+p_u} > \frac{h-l}{2(h-lh)} = \frac{m_k(h-l)}{2m_k(h-lh)} \underset{n\to\infty}{\simeq} \frac{m_k(h-l)+m_\tau l}{2m_k(h-lh)+2m_\tau lh} \Rightarrow \lambda_x > \lambda_y. \tag{61}$$

Then, we have

$$\Pr(\Psi_{\sigma_k} - \Psi_{\sigma_0} \leq 0) \tag{62}$$

$$\underset{stochastically}{\leq} \Pr(X - Y \leq 0) \tag{63}$$

$$\leq 2\exp(-\frac{(\lambda_x-\lambda_y)^2}{8(\lambda_x+\lambda_y)}) \tag{64}$$

$$\leq 2\exp(-\frac{(2l(p_a+p_u-2hp_ap_u)(m_k-m_\tau)-2h(p_a+p_u-2p_ap_u)m_k)^2}{8(2l(p_a+p_u-2hp_ap_u)(m_k-m_\tau)+2h(p_a+p_u-2p_ap_u)m_k)}) \tag{65}$$

$$\underset{n\to\infty}{\simeq} 2\exp(-\frac{(l(p_a+p_u-2hp_ap_u)-h(p_a+p_u-2p_ap_u))^2m_k}{4(l(p_a+p_u-2hp_ap_u)+h(p_a+p_u-2p_ap_u))}) \tag{66}$$

$$= 2\exp(-f_{p_a,p_u}m_k) \tag{67}$$

$$= 2\exp(-\Omega(\frac{2\ln n+1}{kn})m_k) \tag{68}$$

$$\underset{n\to\infty}{\leq} 2\exp(-2\ln n - 1) \tag{69}$$

$$\leq \frac{1}{n^2}. \tag{70}$$

Based on the *Borel-Cantelli Lemma*, it is *a.a.s.* that $\Psi_{\sigma_k} \geq \Psi_{\sigma_0}$ for $2 \leq k \leq n$, i.e., $\Pr(\Psi_\sigma \geq \Psi_{\sigma_0}) \to 1$ for any $\sigma \neq \sigma_0$. □

Again, to guarantee the *uniqueness* of the perfect DA scheme σ_0 to cause the least DE when $p_a \neq p_u$, we quantify the conditions on p_a, p_u, l, and h as follows.

Theorem 4. *When* $g_{p_a,p_u} > \frac{h-l}{2(h-lh)}$ *and* $f_{p_a,p_u} = \Omega(\frac{(k+3)\ln n+1}{kn})$, *where* $2 \leq k \leq n$, *it is* a.a.s. *that there exists no DA scheme* σ *such that* $\sigma \neq \sigma_0$ *and* $\Psi_\sigma \leq \Psi_{\sigma_0}$.

Proof Sketch: This theorem can be proved by employing similar techniques as in Theorem 2. Let $\mathbf{E}$ be the event that *there exists at least one DA scheme* $\sigma \neq \sigma_0$ *such that* $\Psi_\sigma \leq \Psi_{\sigma_0}$. From the proof of Theorem 3 and considering that $f_{p_a,p_u} = \Omega(\frac{(k+3)\ln n+1}{kn})$ for $2 \leq k \leq n$, we have

$$\Pr(\mathbf{E}) \leq \sum_{k=2}^{n} \varrho_k \cdot \Pr(\Psi_{\sigma_k} \leq \Psi_{\sigma_0}) \tag{71}$$

$$\leq \sum_{k=2}^{n} n^k \cdot 2\exp(-f_{p_a,p_u} m_k) \tag{72}$$

$$= \sum_{k=2}^{n} 2\exp(k\ln n - \Omega(\frac{(k+3)\ln n+1}{kn}) m_k) \tag{73}$$

$$\underset{n\to\infty}{\leq} \sum_{k=2}^{n} 2\exp(k\ln n - (k+3)\ln n - 1) \tag{74}$$

$$\leq \sum_{k=2}^{n} \frac{1}{n^3} \tag{75}$$

$$\leq \frac{1}{n^2}. \tag{76}$$

Then, according to the *Borel-Cantelli Lemma*, $\Pr(\mathbf{E}) \to 0$, i.e., it is *a.a.s.* that there exists no DA scheme σ such that $\sigma \neq \sigma_0$ and $\Psi_\sigma \leq \Psi_{\sigma_0}$. □

From Theorem 4, to guarantee the uniqueness of inducing the least DE of σ_0, which is a stronger conclusion compared with that in Theorem 3, the condition on g_{p_a,p_u} is the same as in Theorem 3 and the condition on f_{p_a,p_u} has an increase of $\Theta(k)$. Furthermore, Theorem 4 quantifies the conditions under which the anonymized graph data can be mathematically perfectly de-anonymized when $p_a \neq p_u$.

3.3.4 $(1-\epsilon)$-Perfect De-anonymization

In the aforementioned subsection, the conditions on perfect DA are quantified. Now, we study the conditions on $(1-\epsilon)$-*perfect DA*. Formally, we define a $(1-\epsilon)$-*perfect DA*, denoted by σ^{ϵ}, as a DA scheme under which at most $\epsilon|V^a| = \epsilon n$ users are tolerated to be incorrectly (unsuccessfully) de-anonymized, where $0 \leq \epsilon \leq 1$. Under the $(1-\epsilon)$-perfect DA assumption, any σ_k is proper as long as $k \leq \epsilon n$, i.e., we take it as a satisfiable DA solution. Theoretically, the conditions on $(1-\epsilon)$-perfect DA are quantified in Theorem 5. Note that, when we quantify the conditions for $(1-\epsilon)$-perfect DA, we do not distinguish σ_0 and σ_k with $k \leq \epsilon n$, since they are all proper solutions. Hence, as in the scenario of perfect DA, our quantification takes σ_0 as the reference point.

Theorem 5. *(i) When* $p_a = p_u = \wp$, $\wp > \frac{h-l}{h-hl}$, *and* $f_{\wp} = \Omega(\frac{2\ln n+1}{\epsilon n^2})$, $\Pr(\Psi_{\sigma_k} \geq \Psi_{\sigma_0})$ *for any* σ_k *with* $k > \epsilon n$; *(ii) When* $p_a \neq p_u$, $g_{p_a,p_u} > \frac{h-l}{2(h-lh)}$, *and* $f_{p_a,p_u} = \Omega(\frac{2\ln n+1}{\epsilon n^2})$, $\Pr(\Psi_{\sigma_k} \geq \Psi_{\sigma_0})$ *for any* σ_k *with* $k > \epsilon n$.

Proof Sketch: Since the DA schemes σ_k with $k \leq \epsilon n$ are satisfiable solutions under the $(1-\epsilon)$-perfect DA assumption, we only have to consider σ_k with $k > \epsilon n$.

(i) We first prove the case when $p_a = p_u\wp$. From the proof of Theorem 1, we know that $\Pr(\Psi_{\sigma_k} - \Psi_{\sigma_0} \leq 0) \underset{stochastically}{\leq} \Pr(X - Y \leq 0)$ for $2 \leq k \leq n$, where $X \sim \mathbf{B}(\sum_{e_{i,j}\in\mathcal{E}_k\setminus\mathcal{E}_\tau} 2, p_{i,j}\cdot\wp(1-p_{\sigma_k(i),\sigma_k(j)}\wp)+p_{\sigma_k(i),\sigma_k(j)}\cdot\wp(1-p_{i,j}\wp))$ and $Y \sim \mathbf{B}(\sum_{e_{i,j}\in\mathcal{E}_k} 2, 2p_{i,j}\cdot$

$\wp(1-\wp))$. Now, considering $\wp > \frac{h-l}{h-hl}$, $f_\wp = \Omega(\frac{2\ln n+1}{\epsilon n^2})$, and $k > \epsilon n$, we have

$$\Pr(\Psi_{\sigma_k} - \Psi_{\sigma_0} \leq 0) \leq \Pr(X - Y \leq 0) \tag{77}$$

$$\leq 2\exp(-\frac{(\lambda_x - \lambda_y)^2}{8(\lambda_x + \lambda_y)}) \tag{78}$$

$$\underset{n \to \infty}{\leq} 2\exp(-f_\wp m_k) \tag{79}$$

$$= 2\exp(-\Omega(\frac{2\ln n + 1}{\epsilon n^2}) \cdot \Omega(\epsilon n^2)) \tag{80}$$

$$\leq 2\exp(-2\ln n - 1) \tag{81}$$

$$\leq \frac{1}{n^2}. \tag{82}$$

Consequently, according to the *Borel-Cantelli Lemma*, we have $\Pr(\Psi_{\sigma_k} - \Psi_{\sigma_0} \leq 0) \to 0$ for $k > \epsilon n$ when $p_a = p_u = \wp$.

(ii) We now prove the case when $p_a \neq p_u$. From Theorem 3, when $g_{p_a,p_u} > \frac{h-l}{2(h-lh)}$, we have $\Pr(\Psi_{\sigma_k} - \Psi_{\sigma_0} \leq 0) \leq 2\exp(-f_{p_a,p_u} m_k)$ for $2 \leq k \leq n$. Considering that $k > \epsilon n$ and $f_{p_a,p_u} = \Omega(\frac{2\ln n+1}{\epsilon n^2})$, we have $\Pr(\Psi_{\sigma_k} - \Psi_{\sigma_0} \leq 0) \leq 2\exp(-\Omega(\frac{2\ln n+1}{\epsilon n^2})\Omega(\epsilon n^2)) \leq \frac{1}{n^2}$. Hence, it is *a.a.s.* that $\Psi_{\sigma_k} \geq \Psi_{\sigma_0}$ for σ_k with $k > \epsilon n$ when $p_a \neq p_u$. □

From Theorem 5, we can see that (i) for any DA scheme σ_k, if it has more than ϵn incorrect mappings, with probability 1, it will cause more DE than σ_0. On the other hand, if σ_k is a $(1-\epsilon)$-perfect DA scheme, i.e., $k \leq \epsilon n$, we cannot *a.a.s.* distinguish σ_k and σ_0 based on DE under the quantified conditions; (ii) compared with the quantifications in Theorems 1 and 3, the conditions on $f_\wp$ and f_{p_a,p_u} change from $\Omega(\frac{\ln n}{kn})$ to $\Omega(\frac{\ln n}{n^2})$ explicitly, which implies a relaxation of the condition on $f_\wp$ and f_{p_a,p_u}. This relaxation comes from the toleration of ϵn incorrect user mappings. As in the scenario of perfect DA, stronger conditions can be quantified to guarantee $(1-\epsilon)$-perfect DA schemes causing the least DE. The quantification is shown in Theorem 6, which can be proven by employing similar techniques as in Theorems 2 and 4. Therefore, we omit the detailed proof here. From Theorem 6, we can see that even ϵn matching errors are tolerated, the conditions on $\wp$ and g_{p_a,p_u} stay the same

while the conditions on $f_{\wp}$ and f_{p_a,p_u} only have some constant relaxation compared with the perfect DA scenario.

Theorem 6. *(i) When* $p_a = p_u = \wp$, $\wp > \frac{h-l}{h-hl}$, *and* $f_{\wp} = \Omega(\frac{(\epsilon n+3)\ln n+1}{\epsilon n^2})$, *it is* a.a.s. *that there exists no* σ_k *such that* $k > \epsilon n$ *and* $\Psi_{\sigma_k} \leq \Psi_{\sigma_0}$; *(ii) When* $p_a \neq p_u$, $g_{p_a,p_u} > \frac{h-l}{2(h-lh)}$, *and* $f_{p_a,p_u} = \Omega(\frac{(\epsilon n+3)\ln n+1}{\epsilon n^2})$, *it is* a.a.s. *that there exists no* σ_k *such that* $k > \epsilon n$ *and* $\Psi_{\sigma_k} \leq \Psi_{\sigma_0}$.

3.4 Evaluation

According to our quantification, even without semantic/contextual priori knowledge, anonymized graph data can be de-anonymized perfectly or $(1-\epsilon)$-perfectly when certain structural conditions are satisfied. In this section, we conduct comprehensive evaluations of our DA quantification on 26 real world graph datasets[7].

3.4.1 Evaluation Setup

During the quantification, $p_{i,j}$ is an important parameter although we quantify the conditions in laconic expressions in terms of its bounds l and h. However, it is difficult to accurately determine $p_{i,j}$ in practice [107, 113, 115]. Fortunately, it is not necessary to know the exact $p_{i,j}$ to numerically evaluate our DA quantification. Actually, according to our derivation, we only have to determine the statistical *expectation value* of $p_{i,j}$, denoted by $\mathbb{E}(p_{i,j})$. For a dataset with degree sequence $\mathbf{D}$, define $p_{\mathbf{D}} = \mathbb{E}(p_{i,j})$. Then, it is statistically reasonable (especially for large datasets) to use the *graph density* $\rho = \frac{2m}{n(n-1)}$ to approximate $p_{\mathbf{D}}$, i.e., $p_{\mathbf{D}} \simeq \rho$ [107, 113]. On the other hand, we focus on demonstrating the statistical behavior of our perfect/$(1-\epsilon)$-perfect DA quantification. Therefore, we use ρ to approximate $p_{\mathbf{D}}$ in our evaluation. Furthermore, for the convenience of evaluation, we evaluate the quantification in the scenario

[7]We conduct more evaluations on 60+ real world datasets. Here, partial of the results on 26 representative datasets are shown. Complete results and source codes are available up to request.

of $p_a = p_u = \wp$. This does not limit our evaluation since it is straightforward to extend to the $p_a \neq p_u$ scenario (actually, both scenarios exhibit similar behaviors, which can also be seen in the quantification).

Let

$$f_{\mathbf{D}} = \frac{p_{\mathbf{D}}\wp(\wp - p_{\mathbf{D}}\wp)^2}{2(2 - p_{\mathbf{D}}\wp - \wp)}. \tag{83}$$

Then, we have the following conclusions, which can be proven by similar techniques as in Theorems 1, 2, 5, and 6 from the statistical perspective.

Theorem 7. *For perfect DA, (i) when* $\wp > \frac{k}{(1-p_{\mathbf{D}})(2kn-k^2)+(2p_{\mathbf{D}}-1)k}$ *and* $f_{\mathbf{D}} = \Omega(\frac{4\ln n+2}{2kn-k^2-k})$, $\Pr(\Psi_\sigma \geq \Psi_{\sigma_0}) \to 1$ *for any* $\sigma \neq \sigma_0$*; (ii) when* $\wp > \frac{k}{(1-p_{\mathbf{D}})(2kn-k^2)+(2p_{\mathbf{D}}-1)k}$ *and* $f_{\mathbf{D}} = \Omega(\frac{2(k+3)\ln n+2}{2kn-k^2-k})$*, it is* a.a.s. *that there exists no* σ *such that* $\sigma \neq \sigma_0$ *and* $\Psi_\sigma \leq \Psi_{\sigma_0}$.

Theorem 8. *For* $(1-\epsilon)$*-perfect DA, (i) when* $\wp > \frac{k}{(1-p_{\mathbf{D}})(2kn-k^2)+(2p_{\mathbf{D}}-1)k}$ *and* $f_{\mathbf{D}} = \Omega(\frac{\ln n}{\epsilon n^2})$, $\Pr(\Psi_{\sigma_k} \geq \Psi_{\sigma_0}) \to 1$ *for any* σ_k *with* $k > \epsilon n$*; (ii) when* $\wp > \frac{k}{(1-p_{\mathbf{D}})(2kn-k^2)+(2p_{\mathbf{D}}-1)k}$ *and* $f_{\mathbf{D}} = \Omega(\frac{\ln n}{n})$*, it is* a.a.s. *that there exists no* σ_k *such that* $k > \epsilon n$ *and* $\Psi_\sigma \leq \Psi_{\sigma_0}$.

Now, based on Theorems 7 and 8, we evaluate our quantification on perfect and $(1-\epsilon)$-perfect DA.

3.4.2 Datasets

We evaluate our quantification on 26 datasets from multiple domains, including Social Network (SN) data, Location based Mobility traces and SN (LMSN) data, Collaboration Network (ColN) data, communication network (Email, WikiTalk) data, Autonomous Systems (AS) graph data, and Peer-to-Peer (P2P) network graph data [16, 49, 115, 127]. In Table 3, we show some statistics on the employed datasets, where $\overline{d}$ represents the *average degree* of n nodes and $p(i)$ indicates the *percentage* of nodes with degree of i or less in the corresponding dataset.

Table 3: Data statistics.

Name	Type	n	m	ρ	$\overline{d}$	$p(1)$	$p(5)$
Google+	SN	4.7M	90.8M	8.24E-6	38.7	.054	.273
Twitter	SN	.5M	14.9M	1.20E-4	54.8	.053	.198
LiveJournal	SN	4.8M	69M	3.70E-6	17.9	.210	.505
Facebook	SN	4K	88K	1.08E-2	43.7	.019	.113
YouTube	SN	1.1M	3M	4.64E-6	5.3	.531	.855
Orkut	SN	3.1M	117.2M	2.48E-5	76.3	.022	.073
Slashdot	SN	82.2K	1M	1.73E-4	14.2	.022	.593
Pokec	SN	1.6M	30.6M	1.67E-5	27.3	.100	.307
Infocom	LMSN	73	212	8.07E-2	5.8	.068	.493
Smallblue	LMSN	120	375	5.25E-2	6.3	.133	.625
Brightkite	LMSN	58K	.2M	1.32E-4	7.5	.354	.718
Gowalla	LMSN	.2M	1M	4.92E-5	9.7	.252	.645
HepPh	ColN	12K	.2M	1.87E-3	21.0	.100	.500
AstroPh	ColN	18.8K	.4M	1.23E-3	22.0	.053	.337
CondMat	ColN	23.1K	.2M	4.00E-4	8.6	.078	.518
DBLP	ColN	.3M	1.1M	2.09E-5	6.6	.136	.670
Enron	Email	36.7K	.2M	3.19E-4	10.7	.281	.679
EuAll	Email	.3M	.4M	1.35E-5	3.0	.837	.973
Wiki	WikiTalk	2.4M	5M	1.63E-6	3.9	.738	.962
AS733	AS	6.5K	13.9K	6.63E-4	4.3	.355	.896
Oregon	AS	11.5K	32.7K	4.98E-4	5.7	.289	.876
Caida	AS	26.5K	53.4K	1.52E-4	4.0	.375	.924
Skitter	AS	1.7M	11.1M	7.73E-6	13.1	.128	.554
Gnutella3	P2P	26.5K	65.4K	1.86E-4	4.9	.413	.710
Gnutella4	P2P	36.7K	88.3K	1.32E-4	4.8	.448	.718
Gnutella5	P2P	62.6K	.1M	7.56E-5	4.7	.458	.725

Due to space limitations, we briefly introduce the datasets as follows. Detailed descriptions can be found in [16, 49, 115, 127].

• **SN**. We employed 8 SN datasets in our evaluation as shown in Table 3. Google+ is a SN developed by Google indicating the "circle" relationships (e.g., friends, families, colleagues.) among people [49]. Twitter is a SN that enables users to send and read "tweets" [16]. LiveJournal is a SN that allows members to maintain journals, blogs, etc. [16]. Facebook is a SN where users are connected by "friendships" [16]. In the YouTube and Orkut SNs, users form "friendships" and create groups where other users can join [16]. Slashdot is a SN for sharing and maintaining technology-related news [16]. Pokec is also a "friendship" based SN [16].

• **LMSN**. Infocom consists of a Bluetooth contact trace and a coauthor network of Infocom 2006 conference attendees [127]. Smallblue consists of an *instant messenger* contact trace and a Facebook SN of the employees of a company [127]. Both Brightkite and Gowalla are consisting of a SN and a check-in trace of the SN users [16, 115].

• **ColN**. HepPh, AstroPh, and CondMat are three collaboration networks from arXiv in the areas of *High Energy Physics-Phenomenology*, *Astro Physics*, and *Condense Matter Physics*, respectively [16]. DBLP is a collaboration network of researchers mainly in *Computer Science* [16].

• **Email and WikiTalk**. Enron and EuAll are two email communication networks [16]. WikiTalk is a network containing the discussion relationships among a group of users on Wikipedia [16].

• **AS**. AS733, Oregon, Caida, and Skitter are four AS graphs at different locations [16].

• **P2P**. Gnutella3, Gnutella4, and Gnutella5 are three P2P network graphs where nodes represent hosts in Gnutella and edges are connections between hosts [16].

Before evaluating our quantification, we preprocess the datasets as follows. First, we remove *isolated* users (or nodes) from a dataset if present (most of the datasets

do not have isolated users). This is intuitively reasonable since isolated users carry no structural information. Second, we do not consider the direction information of the directed data, i.e., all the datasets are represented by undirected graphs. This is because our network model is an undirected graph. Even direction takes some extra auxiliary information [104], we do not consider it in this chapter and would include it in the future. More importantly, our quantification demonstrates that undirected structure information is powerful enough to de-anonymize graph data, which can also be seen in our following evaluation.

3.4.3 Evaluation on Perfect De-anonymization Quantification

For each of the datasets considered, we represent it as graph G. Given $\wp$, G^a and G^u can be projected from G according two independent edge/relationship projection processes. Furthermore, the conditions in Theorems 7 and 8 are quantified in the sense of n being a large number. Therefore, in the evaluation of perfect/$(1-\epsilon)$-perfect DA quantification, we also derive an extra condition on the lower bound on n, denoted by $\Omega(n)$. Then, based on Theorem 7, the conditions on $(\Omega(f_{\mathbf{D}}), \Omega(n))$ for perfect DA under different projection probabilities $\wp$ are shown in Table 4.

From Table 4, we have the following observations.

• When $\wp$ increases, $\Omega(f_{\mathbf{D}})$ shows an increasing trend. For instance, $\Omega(f_{\mathbf{D}})$ is increased from 6.5E-8 when $\wp = .3$ to 2.7E-6 when $\wp = .9$, which implies the condition on $f_{\mathbf{D}}$ becomes stronger. This is consistent with our quantification since $f_{\mathbf{D}}$ is an *increasing function* on $\wp$ given $p_{\mathbf{D}}$. On the other hand, we find that although $\Omega(f_{\mathbf{D}})$ increases for large $\wp$, it still keeps relatively loose bounds, i.e., $f_{\mathbf{D}}$ is easily be satisfied. For example, when $\wp = .9$, the condition on $\Omega(f_{\mathbf{D}})$ is 2.7E-6 for Google+ (a large scale dataset) and 1.6E-5 for Gowalla (a medium scale dataset).

• When $\wp$ increases, $\Omega(n)$ decreases. For instance, $\Omega(n)$ is decreased from 1.7E7

Table 4: Evaluation of $(\Omega(f_{\mathbf{D}}), \Omega(n))$ in perfect DA.

Dataset	n	$\wp = .3$	$\wp = .4$	$\wp = .5$	$\wp = .6$	$\wp = .7$	$\wp = .8$	$\wp = .9$
Google+	4.7E6	(6.5E-8, 3.0E8)	(1.6E-7, 1.1E8)	(3.4E-7, 5.2E7)	(6.4E-7, 2.7E7)	(1.1E-6, 1.5E7)	(1.8E-6, 9.1E6)	(2.7E-6, 5.7E6)
Twitter	4.6E5	(9.5E-7, 1.7E7)	(2.4E-6, 6.5E6)	(5.0E-6, 3.0E6)	(9.3E-6, 1.5E6)	(1.6E-5, 8.6E5)	(2.6E-5, 5.1E5)	(4.0E-5, 3.2E5)
LiveJournal	4.8E6	(2.9E-8, 6.9E8)	(7.4E-8, 2.6E8)	(1.5E-7, 1.2E8)	(2.9E-7, 6.3E7)	(4.9E-7, 3.6E7)	(7.9E-7, 2.1E7)	(1.2E-6, 1.3E7)
Facebook	4.0E3	(8.4E-5, 1.4E5)	(2.1E-4, 5.1E4)	(4.4E-4, 2.3E4)	(8.2E-4, 1.1E4)	(1.4E-3, 6.2E3)	(2.3E-3, 3.6E3)	(3.5E-3, 2.2E3)
YouTube	1.1E6	(3.7E-8, 5.5E8)	(9.3E-8, 2.1E8)	(1.9E-7, 9.5E7)	(3.6E-7, 5.0E7)	(6.1E-7, 2.8E7)	(9.9E-7, 1.7E7)	(1.5E-6, 1.1E7)
Orkut	3.1E6	(2.0E-7, 9.3E7)	(5.0E-7, 3.5E7)	(1.0E-6, 1.6E7)	(1.9E-6, 8.3E6)	(3.3E-6, 4.7E6)	(5.3E-6, 2.8E6)	(8.2E-6, 1.7E6)
Slashdot	8.2E4	(1.4E-6, 1.2E7)	(3.5E-6, 4.4E6)	(7.2E-6, 2.0E6)	(1.3E-5, 1.0E6)	(2.3E-5, 5.8E5)	(3.7E-5, 3.5E5)	(5.7E-5, 2.1E5)
Pokec	1.6E6	(1.3E-7, 1.4E8)	(3.3E-7, 5.3E7)	(7.0E-7, 2.4E7)	(1.3E-6, 1.3E7)	(2.2E-6, 7.2E6)	(3.6E-6, 4.3E6)	(5.5E-6, 2.7E6)
Infocom	7.3E1	(5.5E-4, 1.8E4)	(1.4E-3, 6.4E3)	(2.9E-3, 2.7E3)	(5.4E-3, 1.4E3)	(9.4E-3, 7.8E2)	(1.5E-2, 3.9E2)	(2.4E-2, 2.5E2)
Smallblue	1.2E2	(3.8E-4, 2.7E4)	(9.6E-4, 9.7E3)	(2.0E-3, 4.2E3)	(3.7E-3, 2.1E3)	(6.4E-3, 1.2E3)	(1.0E-2, 6.8E2)	(1.6E-2, 4.4E2)
Brightkite	5.7E4	(1.1E-6, 1.6E7)	(2.6E-6, 5.9E6)	(5.5E-6, 2.7E6)	(1.0E-5, 1.4E6)	(1.7E-5, 7.8E5)	(2.8E-5, 4.6E5)	(4.4E-5, 2.9E5)
Gowalla	2.0E5	(3.9E-7, 4.5E7)	(9.8E-7, 1.7E7)	(2.0E-6, 7.7E6)	(3.8E-6, 4.0E6)	(6.5E-6, 2.3E6)	(1.0E-5, 1.3E6)	(1.6E-5, 8.4E5)
HepPh	1.2E4	(1.5E-5, 9.3E5)	(3.7E-5, 3.4E5)	(7.8E-5, 1.5E5)	(1.4E-4, 7.8E4)	(2.5E-4, 4.3E4)	(4.0E-4, 2.6E4)	(6.2E-4, 1.6E4)
AstroPh	1.8E4	(9.7E-6, 1.5E6)	(2.5E-5, 5.4E5)	(5.1E-5, 2.4E5)	(9.5E-5, 1.2E5)	(1.6E-4, 6.9E4)	(2.6E-4, 4.1E4)	(4.1E-4, 2.5E4)
CondMat	2.1E4	(3.2E-6, 4.8E6)	(8.0E-6, 1.8E6)	(1.7E-5, 8.2E5)	(3.1E-5, 4.2E5)	(5.3E-5, 2.3E5)	(8.5E-5, 1.4E5)	(1.3E-4, 8.6E4)
DBLP	3.2E5	(1.7E-7, 1.1E8)	(4.2E-7, 4.2E7)	(8.7E-7, 1.9E7)	(1.6E-6, 1.0E7)	(2.8E-6, 5.6E6)	(4.5E-6, 3.4E6)	(6.9E-6, 2.1E6)
Enron	3.4E4	(2.5E-6, 6.2E6)	(6.4E-6, 2.3E6)	(1.3E-5, 1.0E6)	(2.5E-5, 5.4E5)	(4.2E-5, 3.0E5)	(6.8E-5, 1.8E5)	(1.1E-4, 1.1E5)
EuAll	2.2E5	(1.1E-7, 1.8E8)	(2.7E-7, 6.7E7)	(5.6E-7, 3.1E7)	(1.0E-6, 1.6E7)	(1.8E-6, 9.0E6)	(2.9E-6, 5.4E6)	(4.5E-6, 3.4E6)
Wiki	2.4E6	(1.3E-8, 1.6E9)	(3.3E-8, 6.2E8)	(6.8E-8, 2.9E8)	(1.3E-7, 1.5E8)	(2.2E-7, 8.5E7)	(3.5E-7, 5.1E7)	(5.4E-7, 3.2E7)
AS733	6.5E3	(5.3E-6, 2.8E6)	(1.3E-5, 1.0E6)	(2.8E-5, 4.7E5)	(5.1E-5, 2.4E5)	(8.7E-5, 1.4E5)	(1.4E-4, 8.0E4)	(2.2E-4, 4.9E4)
Oregon	1.1E4	(4.0E-6, 3.8E6)	(1.0E-5, 1.4E6)	(2.1E-5, 6.4E5)	(3.8E-5, 3.3E5)	(6.6E-5, 1.8E5)	(1.1E-4, 1.1E5)	(1.7E-4, 6.7E4)
Caida	2.6E4	(1.2E-6, 1.4E7)	(3.0E-6, 5.1E6)	(6.3E-6, 2.3E6)	(1.2E-5, 1.2E6)	(2.0E-5, 6.7E5)	(3.2E-5, 4.0E5)	(5.0E-5, 2.5E5)
Skitter	1.7E6	(6.1E-8, 3.2E8)	(1.5E-7, 1.2E8)	(3.2E-7, 5.5E7)	(6.0E-7, 2.9E7)	(1.0E-6, 1.6E7)	(1.6E-6, 9.8E6)	(2.6E-6, 6.1E6)
Gnutella3	2.6E4	(1.5E-6, 1.1E7)	(3.7E-6, 4.1E6)	(7.8E-6, 1.9E6)	(1.4E-5, 9.6E5)	(2.5E-5, 5.4E5)	(4.0E-5, 3.2E5)	(6.2E-5, 2.0E5)
Gnutella4	3.7E4	(1.0E-6, 1.6E7)	(2.6E-6, 5.9E6)	(5.5E-6, 2.7E6)	(1.0E-5, 1.4E6)	(1.7E-5, 7.8E5)	(2.8E-5, 4.7E5)	(4.4E-5, 2.9E5)
Gnutella5	6.3E4	(6.0E-7, 2.9E7)	(1.5E-6, 1.1E7)	(3.1E-6, 4.9E6)	(5.8E-6, 2.5E6)	(1.0E-5, 1.4E6)	(1.6E-5, 8.5E5)	(2.5E-5, 5.3E5)

when $\wp = .3$ to 3.2E5 when $\wp = .9$ for Twitter. This is because a large $\wp$ implies that G^a is topologically more similar to G^u. Thus, a weaker condition on $\Omega(n)$ is sufficient to enable a perfect DA scheme *a.a.s.* inducing the least DE.

• For datasets with similar graph densities, e.g., Google+ (ρ = 8.24E-6) and Skitter (ρ = 7.73E-6), the conditions on $(\Omega(f_{\mathbf{D}}), \Omega(n))$ are also similar for perfect DA, which is consistent with our theoretical quantification. This comes from the similarity of their statistical $p_{\mathbf{D}}$. For perfect DA on datasets with different graph densities (with similar or different sizes), e.g., HepPh (n = 1.2E4, ρ = 1.87E-3) and Oregon (n = 1.15E4, ρ = 4.98E-4), Facebook (n = 4.0E3, ρ = 1.08E-2) and Twitter (n = 4.6E5, ρ = 1.2E-4), dense datasets require a stronger condition on $f_{\mathbf{D}}$ while a weaker condition on $\Omega(n)$ given $\wp$, which is also consistent with our quantification. A stronger condition requirement on $f_{\mathbf{D}}$ is because of that $f_{\mathbf{D}}$ is an increasing function on $p_{\mathcal{D}} \simeq \rho \in (0, 0.5]$ given $\wp$ and all the considering datasets have $\rho \leq 0.5$. A looser bound on $\Omega(n)$ comes from the fact that more structural information can be projected to G^a and G^u in dense datasets.

• From Table 4, some datasets can be perfectly de-anonymized under some conditions. For instance, Orkut and Facebook are *a.a.s.* can be perfectly de-anonymized when $\wp \geq \Omega(.8)$, and Twitter is *a.a.s.* can be perfectly de-anonymized when $\wp \geq \Omega(.9)$. The perfect DA is due to their good structural characteristics, e.g., high average degree (from Table 3, the average degree $\overline{d}$ is 76.3 for Orkut, 54.8 for Twitter, and 43.7 for Facebook), small percentage of nodes with a low degree ($p(1)$ is 2.2% for Orkut, 5.3% for Twitter, and 5.4% for Facebook).

3.4.4 Evaluation on $(1-\epsilon)$-Perfect De-anonymization Quantification

Based on our quantification, the percentage of successfully de-anonymized users by any $(1-\epsilon)$-perfect DA scheme is at least $1-\epsilon$. Given $\wp$ varied from .3 to .95, we evaluate the minimum number of users in the 26 datasets considered that can be

successfully de-anonymized with probability 1 in terms of our quantification, i.e., the lower bound of $1-\epsilon$, $(\Omega(1-\epsilon))$, and the results are shown in Table 5.

From Table 5, we make some important observations and comments as follows.

• When $\wp$ increases, more users can be de-anonymized for every dataset as expected. For example, when $\wp = .5$, it is *a.a.s.* at least 29.7% of the users in Google+ can be successfully de-anonymized; when $\wp$ is increased to .8, at least 72.5% of the users in Google+ can be successfully de-anonymized; when $\wp = .95$ all the users in Google+ can *a.a.s.* be successfully de-anonymized. From Table 5, similar DA phenomena applied to all the datasets, which is consistent with our quantification. The reason is straightforward. When $\wp$ increases, more edges/relationships appear in both G^a and G^u (the expected number of common edges is $m\wp^2$). Thus, the structural similarity between G^a and G^u is increased followed by more users can statistically be successfully de-anonymized with probability 1.

• Most of the existing graph datasets, including SN data, LMSN data, Email and Wiki data, AS data, P2P data, etc., are *a.a.s.* de-anonymizable completely or at least partially just based on the topological information. For instance, Facebook and Orkut datasets can be completely de-anonymized when $\wp = .8$, Twitter can be completely de-anonymized when $\wp = .85$, and Google+ can be completely de-anonymized when $\wp = .95$. Even a dataset cannot be completely de-anonymized, it may be de-anonymizable partially in a large-scale. For example, when $\wp = .9$, at least 60.9%, 48.9%, and 85.7% of the users in LiveJournal, Gowalla, and AstroPh can be successfully de-anonymized, respectively. This fact is consistent with our quantification as well as the intuition that structure itself can be used to de-anonymize data.

• An interesting observation is that the DA results on two datasets with similar graph densities may be very different in practice. From Table 4, for two datasets with

Table 5: Evaluation of $\Omega(1-\epsilon)$ in $(1-\epsilon)$-perfect DA.

Dataset	$\wp = .3$	$\wp = .35$	$\wp = .4$	$\wp = .45$	$\wp = .5$	$\wp = .55$	$\wp = .6$	$\wp = .65$	$\wp = .7$	$\wp = .75$	$\wp = .8$	$\wp = .85$	$\wp = .9$	$\wp = .95$
Google+	11.7%	15.5%	19.7%	24.5%	29.7%	35.5%	41.8%	48.7%	56.1%	64.0%	72.5%	81.6%	91.2%	100.0%
Twitter	15.1%	20.2%	26.0%	32.4%	39.4%	47.1%	55.4%	64.3%	73.8%	84.0%	94.7%	100.0%	100.0%	100.0%
LiveJournal	6.6%	9.1%	11.9%	15.2%	18.8%	22.7%	27.1%	31.8%	36.8%	42.3%	48.1%	54.3%	60.9%	68.1%
Facebook	3.7%	12.1%	22.4%	31.0%	39.9%	49.5%	59.6%	70.3%	81.5%	93.2%	100.0%	100.0%	100.0%	100.0%
YouTube	4.0%	5.3%	6.8%	8.4%	10.3%	12.3%	14.5%	16.9%	19.5%	22.4%	25.5%	28.9%	32.5%	36.4%
Orkut	14.2%	19.6%	26.0%	33.3%	41.4%	50.3%	60.0%	70.4%	81.3%	92.7%	100.0%	100.0%	100.0%	100.0%
Slashdot	7.2%	9.8%	12.7%	15.9%	19.5%	23.4%	27.6%	32.2%	37.2%	42.7%	48.6%	54.9%	61.8%	69.3%
Pokec	7.3%	10.4%	14.1%	18.4%	23.2%	28.5%	34.4%	40.7%	47.5%	54.7%	62.4%	70.5%	79.0%	88.1%
Infocom	10.4%	11.5%	12.5%	13.0%	13.9%	14.3%	15.1%	15.5%	15.8%	16.6%	16.9%	17.2%	49.9%	62.2%
Smallblue	8.9%	9.6%	10.3%	10.9%	11.3%	11.8%	12.1%	12.6%	12.9%	13.3%	33.0%	44.6%	54.6%	64.7%
Brightkite	4.7%	6.5%	8.6%	10.9%	13.5%	16.4%	19.6%	23.1%	26.8%	30.9%	35.3%	40.0%	45.1%	50.6%
Gowalla	5.3%	7.2%	9.4%	11.9%	14.7%	17.8%	21.2%	25.0%	29.0%	33.4%	38.2%	43.3%	48.9%	54.8%
HepPh	9.0%	13.2%	17.6%	22.4%	27.6%	33.2%	39.2%	45.7%	52.7%	60.1%	68.1%	76.7%	85.9%	95.7%
AstroPh	7.4%	11.0%	15.3%	20.1%	25.4%	31.2%	37.6%	44.4%	51.7%	59.4%	67.6%	76.4%	85.7%	95.6%
CondMat	3.5%	5.2%	7.2%	9.6%	12.3%	15.3%	18.7%	22.6%	26.8%	31.4%	36.5%	42.1%	48.2%	54.8%
DBLP	3.0%	4.3%	5.8%	7.6%	9.6%	11.8%	14.3%	17.1%	20.2%	23.6%	27.4%	31.5%	36.0%	40.9%
Enron	6.6%	9.0%	11.7%	14.6%	17.9%	21.4%	25.3%	29.5%	34.1%	39.1%	44.5%	50.3%	56.6%	63.4%
EuAll	3.5%	4.5%	5.6%	6.9%	8.3%	9.8%	11.4%	13.3%	15.2%	17.4%	19.6%	22.1%	24.7%	27.6%
Wiki	3.7%	4.8%	6.0%	7.4%	8.9%	10.5%	12.3%	14.2%	16.3%	18.6%	21.1%	23.8%	26.7%	29.8%
AS733	1.3%	4.8%	6.5%	8.3%	10.3%	12.5%	14.9%	17.6%	20.5%	23.8%	27.4%	31.2%	35.5%	40.0%
Oregon	4.6%	6.5%	8.6%	10.8%	13.1%	15.7%	18.5%	21.6%	24.9%	28.6%	32.5%	36.7%	41.3%	46.3%
Caida	3.8%	5.1%	6.5%	8.1%	9.9%	11.8%	14.0%	16.3%	18.8%	21.6%	24.6%	27.8%	31.4%	35.3%
Skitter	6.2%	8.3%	10.6%	13.3%	16.2%	19.5%	23.1%	27.1%	31.4%	36.1%	41.2%	46.7%	52.6%	59.1%
Gnutella3	1.7%	2.6%	3.8%	5.4%	7.2%	9.5%	12.1%	15.2%	18.8%	23.0%	27.3%	31.5%	36.0%	40.6%
Gnutella4	1.8%	2.8%	4.0%	5.5%	7.3%	9.4%	12.0%	15.0%	18.4%	22.5%	26.7%	30.8%	35.1%	39.6%
Gnutella5	1.8%	2.7%	3.9%	5.3%	7.0%	9.1%	11.5%	14.4%	17.7%	21.6%	25.7%	29.7%	33.8%	38.1%

similar graph densities, e.g., Google+ (ρ = 8.24E-6) and Skitter (ρ = 7.73E-6), the theoretical bounds on $(\Omega(f_{\mathbf{D}}), \Omega(n))$ for perfect DA are also similar. However, from Table 5, the DA results of Google+ and Skitter are very different: when $\wp = .6$, the number of de-anonymizable users in Google+ (41.8%) is about twice of that in Skitter (23.1%); while when $\wp = .95$, all the users in Google+ are *a.a.s.* de-anonymizable while the de-anonymizable users in Skitter is only bounded by $\Omega(59.1\%)$. To study the reason of this fact, we need to consider the degree distribution of Google+ and Skitter besides the graph density (as well as $\Omega(f_{\mathbf{D}})$ and $\Omega(n)$). From Table 3, the percentage of low degree users in Skitter ($p(1) = 12.8\%$ and $p(5) = 55.4\%$) is much higher than that in Google+ ($p(1) = 5.4\%$ and $p(5) = 27.3\%$). On the other hand, intuitively, low degree users, especially users with degree of 1, do not have too much distinguishable structural information (this intuition is confirmed by our theoretical quantification on different DEs caused by mismatching high degree users and low degree users), which implies that they are difficult to be de-anonymized based on structural information. Consequently, the existence of a large amount of low degree users in Skitter makes it less de-anonymizable than Google+, which is consistent with our quantification. In summary, from Tables 3 and 5, if a dataset has a high average degree and a small percentage of low degree users, e.g., Orkut, Facebook, Twitter, Google+, it is easier to de-anonymize and a large amount of its users are *a.a.s.* de-anonymizable; otherwise, for datasets with a low average degree and a large percentage of low degree users, e.g., EuAll, Wiki, Caida, they are difficult to be de-anonymized based solely on the structural information.

• Following the above observation, we find that there exists some difference between theory and practice on the dominating factor of DA. Theoretically, the graph density plays as a dominating factor on determining the bound of $(\Omega(f_{\mathbf{D}}), \Omega(n))$ (Table 4). In practice, the degree distribution and the average degree have more impact on the DA results (Table 5). This is mainly because that we study the quantification

Table 6: Evaluation of $\Omega(n)$ in $(1-\epsilon)$-perfect DA.

Dataset	$\wp = .3$				$\wp = .6$				$\wp = .9$			
	$\epsilon = .1$	$\epsilon = .2$	$\epsilon = .3$	$\epsilon = .4$	$\epsilon = .1$	$\epsilon = .2$	$\epsilon = .3$	$\epsilon = .4$	$\epsilon = .1$	$\epsilon = .2$	$\epsilon = .3$	$\epsilon = .4$
Google+	2.4E8	1.9E8	1.4E8	1.1E8	2.2E7	1.7E7	1.3E7	9.5E6	4.6E6	3.6E6	2.7E6	2.3E6
Twitter	1.4E7	1.1E7	8.4E6	6.2E6	1.2E6	9.6E5	7.3E5	5.4E5	2.5E5	2.3E5	2.3E5	2.3E5
LiveJournal	5.6E8	4.4E8	3.4E8	2.5E8	5.1E7	4.0E7	3.0E7	2.2E7	1.1E7	8.4E6	6.4E6	4.7E6
Facebook	1.1E5	9.0E4	6.9E4	5.2E4	9.2E3	7.2E3	5.5E3	4.1E3	2.0E3	2.0E3	2.0E3	2.0E3
YouTube	4.5E8	3.6E8	2.7E8	2.0E8	4.0E7	3.2E7	2.5E7	1.8E7	8.6E6	6.8E6	5.2E6	3.8E6
Orkut	7.5E7	5.9E7	4.6E7	3.5E7	6.7E6	5.3E6	4.1E6	3.1E6	1.5E6	1.5E6	1.5E6	1.5E6
Slashdot	9.7E6	7.7E6	5.9E6	4.4E6	8.5E5	6.7E5	5.2E5	3.8E5	1.7E5	1.4E5	1.1E5	7.7E4
Pokec	1.1E8	8.9E7	6.8E7	5.1E7	1.0E7	8.0E6	6.1E6	4.5E6	2.1E6	1.7E6	1.3E6	9.4E5
Infocom	1.5E4	1.3E4	1.1E4	9.0E3	1.2E3	9.8E2	7.8E2	6.8E2	2.5E2	2.5E2	1.7E2	1.7E2
Smallblue	2.2E4	1.8E4	1.5E4	1.2E4	1.8E3	1.4E3	1.2E3	8.8E2	3.4E2	3.2E2	2.2E2	2.2E2
Brightkite	1.3E7	1.0E7	7.7E6	5.7E6	1.1E6	8.8E5	6.7E5	4.9E5	2.3E5	1.8E5	1.4E5	1.0E5
Gowalla	3.6E7	2.9E7	2.2E7	1.6E7	3.2E6	2.5E6	1.9E6	1.4E6	6.7E5	5.3E5	4.0E5	3.0E5
HepPh	7.4E5	5.8E5	4.4E5	3.2E5	6.2E4	4.9E4	3.7E4	2.7E4	1.2E4	9.7E3	7.3E3	5.6E3
AstroPh	1.2E6	9.2E5	7.0E5	5.2E5	9.9E4	7.8E4	5.9E4	4.4E4	2.0E4	1.6E4	1.2E4	9.0E3
CondMat	3.9E6	3.1E6	2.4E6	1.9E6	3.4E5	2.7E5	2.1E5	1.6E5	6.9E4	5.5E4	4.2E4	3.2E4
DBLP	9.1E7	7.3E7	5.7E7	4.3E7	8.1E6	6.5E6	5.0E6	3.8E6	1.7E6	1.4E6	1.1E6	8.0E5
Enron	5.0E6	3.9E6	3.0E6	2.2E6	4.3E5	3.4E5	2.6E5	1.9E5	8.8E4	6.9E4	5.2E4	3.8E4
EuAll	1.5E8	1.2E8	9.3E7	7.0E7	1.3E7	1.1E7	8.3E6	6.2E6	2.8E6	2.2E6	1.7E6	1.3E6
Wiki	1.3E9	1.1E9	8.4E8	6.3E8	1.2E8	9.9E7	7.7E7	5.7E7	2.6E7	2.1E7	1.6E7	1.2E7
AS733	2.3E6	1.8E6	1.4E6	1.1E6	2.0E5	1.6E5	1.2E5	9.0E4	4.0E4	3.2E4	2.4E4	1.8E4
Oregon	3.1E6	2.5E6	1.9E6	1.4E6	2.7E5	2.1E5	1.6E5	1.2E5	5.5E4	4.3E4	3.3E4	2.4E4
Caida	1.1E7	8.9E6	6.9E6	5.1E6	9.8E5	7.8E5	6.0E5	4.5E5	2.0E5	1.6E5	1.2E5	9.1E4
Skitter	2.6E8	2.0E8	1.6E8	1.2E8	2.3E7	1.8E7	1.4E7	1.0E7	4.9E6	3.9E6	3.0E6	2.2E6
Gnutella3	9.0E6	7.1E6	5.5E6	4.0E6	7.8E5	6.2E5	4.8E5	3.5E5	1.6E5	1.3E5	9.7E4	7.1E4
Gnutella4	1.3E7	1.0E7	8.0E6	5.9E6	1.1E6	9.0E5	6.9E5	5.1E5	2.3E5	1.9E5	1.4E5	1.0E5
Gnutella5	2.3E7	1.9E7	1.4E7	1.1E7	2.1E6	1.6E6	1.3E6	9.3E5	4.3E5	3.4E5	2.6E5	1.9E5

from an asymptotical sense in the theoretical scenario (i.e., $n \to \infty$) and the key parameter $p_{i,j}$ asymptotically converges to graph density ρ, i.e., $\mathbb{E}(p_{i,j}) \underset{n \to \infty}{\simeq} \rho$. On the other hand, when quantifying the percentage of de-anonymizable users for each dataset, the actual degree sequence/distribution $\mathbf{D}$ is used to examine when the DA conditions are satisfied.

We also evaluate the impact of $\wp$ and ϵ on the bound of $\Omega(n)$ in $(1-\epsilon)$-perfect DA (we do not show $\Omega(f_{\mathbf{D}})$ since it depends on $\wp$ and exhibits the same behavior as in the perfect DA). The results are shown in Table 6. From Table 6, we have the following observations.

- When ϵ is fixed, the impact of $\wp$ on $\Omega(n)$ in $(1-\epsilon)$-perfect DA is similar to that in perfect DA, i.e., when $\wp$ increases, $\Omega(n)$ decreases. The reason is also the same as before since a large $\wp$ implies more similarity between G^a and G^u and thus a loose condition on $\Omega(n)$ is sufficient to enable σ_k ($k \leq \epsilon n$) to induce less DE than

$\sigma_{k'}$ $(k' > \epsilon n)$.

• When $\wp$ is fixed, $\Omega(n)$ is also decreasing with the increase of ϵ. For instance, when $\wp = 0.6$, $\Omega(n)$ is decreased from 2.2E7 to 9.5E6 for Google+ when ϵ is increased from .1 to .4. This is because of that when ϵ increases, more DE is tolerated, and thus loose condition is required for $\Omega(n)$ to distinguish σ_k $(k \leq \epsilon n)$ and $\sigma_{k'}$ $(k' > \epsilon n)$, which is consistent with our quantification.

• As in the perfect DA scenario, graph density is an important factor to impact $\Omega(n)$. Datasets with similar graph density, e.g., Google+ and Skitter, exhibits similar requirement on $\Omega(n)$. A dataset with high graph density, e.g., Facebook and HepPh, corresponds to a loose bound on $\Omega(n)$. The reason is also the same as before.

Finally, we also want to evaluate the required bounds on $(\Omega(\wp), \Omega(f_{\mathbf{D}}), \Omega(n))$ in $(1-\epsilon)$-perfect DA. We demonstrate the results in Table 7 and make the following observations.

• Theoretically, the condition on the lower bound of $\wp$ is very loose, e.g., when $\epsilon = .1$, $\Omega(\wp) = 1.1\text{E-}7$ for Google+ and $\Omega(\wp) = 1.7\text{E-}7$ for Orkut, which suggests that $(1-\epsilon)$-perfect DA is implementable in practice. On the other hand, we can also see that the theoretical loose requirement on $\Omega(\wp)$ is at the expense of a strong condition on $\Omega(n)$, e.g., when $\epsilon = .1$, $\Omega(n) = 2.2\text{E}28$ for Google+ and $\Omega(n) = 2.0\text{E}27$ for Orkut. Consequently, to de-anonymize most of existing graph datasets which have sizes of million-level or less, a higher $\wp$ is desired (as we show in Tables 4, 5, and 6).

• From Table 7, we can see that the conditions on $\Omega(f_{\mathbf{D}})$ and $\Omega(n)$ exhibit the same behavior as in perfect DA, i.e., $\Omega(f_{\mathbf{D}})$ increases and $\Omega(n)$ decreases as $\Omega(\wp)$ increases, which is consistent with our quantification. Again, this is because $f_{\mathbf{D}}$ is an increasing function of $\wp$ given $p_{\mathbf{D}}$ and $\Omega(n)$ decreases when more similarity appears between G^a and G^u.

Table 7: Evaluation of $(\Omega(\wp), \Omega(f_{\mathbf{D}}), \Omega(n))$ in $(1-\epsilon)$-perfect DA.

Dataset	$\epsilon = .1$	$\epsilon = .2$	$\epsilon = .3$	$\epsilon = .4$	$\epsilon = .5$
Google+	(1.1E-7, 3.5E-6, 2.2E28)	(1.2E-7, 3.7E-6, 1.8E28)	(1.3E-7, 3.9E-6, 1.6E28)	(1.4E-7, 4.2E-6, 1.3E28)	(1.4E-7, 4.4E-6, 1.1E28)
Twitter	(1.2E-6, 3.1E-5, 1.2E24)	(1.2E-6, 3.2E-5, 9.7E23)	(1.3E-6, 3.4E-5, 8.3E23)	(1.4E-6, 3.6E-5, 6.9E23)	(1.5E-6, 3.8E-5, 5.8E23)
LiveJournal	(1.2E-7, 3.6E-6, 4.7E28)	(1.2E-7, 3.6E-6, 4.7E28)	(1.2E-7, 3.8E-6, 3.8E28)	(1.3E-7, 4.1E-6, 3.0E28)	(1.4E-7, 4.3E-6, 2.7E28)
Facebook	(1.3E-4, 2.2E-3, 5.9E15)	(1.4E-4, 2.3E-3, 5.0E15)	(1.5E-4, 2.5E-3, 4.1E15)	(1.6E-4, 2.6E-3, 3.5E15)	(1.7E-4, 2.8E-3, 2.9E15)
YouTube	(6.0E-7, 1.7E-5, 2.4E26)	(6.0E-7, 1.7E-5, 2.4E26)	(6.0E-7, 1.7E-5, 2.4E26)	(6.0E-7, 1.7E-5, 2.4E26)	(6.0E-7, 1.7E-5, 2.4E26)
Orkut	(1.7E-7, 5.1E-6, 2.0E27)	(1.8E-7, 5.4E-6, 1.7E27)	(1.9E-7, 5.7E-6, 1.4E27)	(2.0E-7, 6.1E-6, 1.2E27)	(2.2E-7, 6.5E-6, 9.8E26)
Slashdot	(7.4E-6, 1.7E-4, 2.8E21)	(7.4E-6, 1.7E-4, 2.8E21)	(7.4E-6, 1.7E-4, 2.8E21)	(8.2E-6, 1.9E-4, 2.1E21)	(8.2E-6, 1.9E-4, 2.1E21)
Pokec	(3.2E-7, 9.2E-6, 4.4E26)	(3.5E-7, 9.9E-6, 3.6E26)	(3.6E-7, 1.0E-5, 3.1E26)	(3.9E-7, 1.1E-5, 2.5E26)	(4.1E-7, 1.2E-5, 2.1E26)
Infocom	(8.6E-3, 8.0E-2, 2.0E09)	(8.6E-3, 8.0E-2, 2.0E09)	(9.1E-3, 8.1E-2, 1.7E09)	(1.0E-2, 8.6E-2, 1.2E09)	(1.1E-2, 8.9E-2, 1.1E09)
Smallblue	(4.7E-3, 5.2E-2, 1.9E10)	(5.1E-3, 5.1E-2, 1.5E10)	(5.3E-3, 5.2E-2, 1.3E10)	(5.8E-3, 5.6E-2, 1.0E10)	(6.4E-3, 6.1E-2, 7.2E09)
Brightkite	(1.1E-5, 2.3E-4, 1.2E21)	(1.1E-5, 2.3E-4, 1.2E21)	(1.1E-5, 2.3E-4, 1.2E21)	(1.2E-5, 2.6E-4, 8.7E20)	(1.2E-5, 2.6E-4, 8.7E20)
Gowalla	(2.9E-6, 7.1E-5, 1.8E23)	(2.9E-6, 7.1E-5, 1.8E23)	(3.2E-6, 7.8E-5, 1.3E23)	(3.2E-6, 7.8E-5, 1.3E23)	(3.4E-6, 8.3E-5, 1.1E23)
HepPh	(4.7E-5, 8.8E-4, 8.5E17)	(5.1E-5, 9.5E-4, 6.7E17)	(5.5E-5, 1.0E-3, 5.3E17)	(5.7E-5, 1.1E-3, 4.6E17)	(6.2E-5, 1.2E-3, 3.7E17)
AstroPh	(3.0E-5, 5.9E-4, 5.2E18)	(3.1E-5, 6.1E-4, 4.6E18)	(3.4E-5, 6.6E-4, 3.7E18)	(3.5E-5, 6.9E-4, 3.1E18)	(3.7E-5, 7.3E-4, 2.6E18)
CondMat	(2.6E-5, 5.2E-4, 2.5E19)	(2.6E-5, 5.2E-4, 2.5E19)	(2.8E-5, 5.6E-4, 2.0E19)	(3.0E-5, 6.0E-4, 1.7E19)	(3.2E-5, 6.3E-4, 1.4E19)
DBLP	(1.7E-6, 4.3E-5, 2.2E24)	(1.9E-6, 4.8E-5, 1.6E24)	(1.9E-6, 4.8E-5, 1.6E24)	(2.1E-6, 5.3E-5, 1.2E24)	(2.2E-6, 5.7E-5, 9.5E23)
Enron	(1.7E-5, 3.6E-4, 1.1E20)	(1.7E-5, 3.6E-4, 1.1E20)	(1.8E-5, 3.8E-4, 9.3E19)	(2.0E-5, 4.2E-4, 7.1E19)	(2.0E-5, 4.2E-4, 7.1E19)
EuAll	(3.8E-6, 9.4E-5, 2.9E23)	(3.8E-6, 9.4E-5, 2.9E23)	(3.8E-6, 9.4E-5, 2.9E23)	(3.8E-6, 9.4E-5, 2.9E23)	(3.8E-6, 9.4E-5, 2.9E23)
Wiki	(3.3E-7, 9.7E-6, 4.3E27)	(3.3E-7, 9.7E-6, 4.3E27)	(3.3E-7, 9.7E-6, 4.3E27)	(3.3E-7, 9.7E-6, 4.3E27)	(3.3E-7, 9.7E-6, 4.3E27)
AS733	(9.4E-5, 1.7E-3, 2.9E17)	(9.4E-5, 1.7E-3, 2.9E17)	(9.4E-5, 1.7E-3, 2.9E17)	(1.1E-4, 2.0E-3, 1.6E17)	(1.1E-4, 2.0E-3, 1.6E17)
Oregon	(5.1E-5, 9.5E-4, 2.6E18)	(5.1E-5, 9.5E-4, 2.6E18)	(6.7E-5, 1.3E-3, 1.1E18)	(6.7E-5, 1.3E-3, 1.1E18)	(6.7E-5, 1.3E-3, 1.1E18)
Caida	(2.3E-5, 4.7E-4, 9.6E19)	(2.3E-5, 4.7E-4, 9.6E19)	(2.3E-5, 4.7E-4, 9.6E19)	(3.1E-5, 6.3E-4, 4.1E19)	(3.1E-5, 6.3E-4, 4.1E19)
Skitter	(3.2E-7, 9.0E-6, 1.0E27)	(3.4E-7, 9.8E-6, 8.0E26)	(3.7E-7, 1.1E-5, 6.5E26)	(3.9E-7, 1.1E-5, 5.4E26)	(4.1E-7, 1.2E-5, 4.7E26)
Gnutella3	(2.4E-5, 4.8E-4, 7.3E19)	(2.4E-5, 4.8E-4, 7.3E19)	(2.4E-5, 4.8E-4, 7.3E19)	(2.4E-5, 4.8E-4, 7.3E19)	(2.6E-5, 5.3E-4, 5.4E19)
Gnutella4	(1.8E-5, 3.7E-4, 2.6E20)	(1.8E-5, 3.7E-4, 2.6E20)	(1.8E-5, 3.7E-4, 2.6E20)	(1.8E-5, 3.7E-4, 2.6E20)	(1.9E-5, 4.1E-4, 2.0E20)
Gnutella5	(1.0E-5, 2.3E-4, 2.3E21)	(1.0E-5, 2.3E-4, 2.3E21)	(1.0E-5, 2.3E-4, 2.3E21)	(1.0E-5, 2.3E-4, 2.3E21)	(1.2E-5, 2.5E-4, 1.7E21)

• From Table 7, we can also see that the impact of graph density on $\Omega(f_{\mathbf{D}})$ and $\Omega(n)$ is also similar to that in the perfect DA scenario.

3.5 Optimization based De-anonymization Practice

In Section 3.3, we comprehensively quantify conditions for perfect DA and $(1-\epsilon)$-perfect DA. Based on our large-scale study on 26 real world datasets in Section 3.4, we find most, if not all, existing graph datasets are de-anonymizable partially or completely (Table 5). Interestingly, our DA quantification leads to a DA scheme, denoted by $\mathfrak{A}^*$, straightforwardly. Basically, $\mathfrak{A}^*$ can be implemented as follows: we can calculate the DE caused by each σ_k $(1 \leq k \leq n!)$ and let σ_0 be the σ_k that induces the least DE. According to the quantification, the σ_0 produced by $\mathfrak{A}^*$ should be the optimum DA scheme. However, $\mathfrak{A}^*$ is computationally infeasible in practice due to its high computational complexity $O(n!)$. In this section, we present a novel relaxed and operational version of $\mathfrak{A}^*$ followed by analyzing its performance theoretically and experimentally on large scale real datasets.

3.5.1 Optimization based De-anonymization

Before proposing our relaxed and computationally feasible version of $\mathfrak{A}^*$, we define some useful *structural features* for $i \in V^a$ or V^u as follows.

• *Degree:* For $i \in V^a$ (resp., V^u), its *degree feature* $f_d(i)$ is its degree in G^a (resp., G^u), i.e., $f_d(i) = |N_i^a|$ (resp., $|N_i^u|$).

• *Neighborhood:* For $i \in V^a$ (resp., V^u), its *neighborhood feature* $\overline{f_n(i)}$ is a β-*dimensional vector* $(d_1^i, d_2^i, \cdots, d_\beta^i)$, where d_k^i $(1 \leq k \leq \beta)$ is the k-th largest degree in $\{|N_j^a||j \in N_i^a\}$ (resp., $\{|N_j^u||j \in N_i^u\}$), i.e., d_k^i is the k-th largest degree of the neighboring users of i. In the case that $|N_i^a| < \beta$ (resp., $|N_i^u| < \beta$), we set $d^i_{|N_i^a|+1} = d^i_{|N_i^a|+2} = \cdots = d^i_\beta = \Delta^a$ (resp., $d^i_{|N_i^u|+1} = d^i_{|N_i^u|+2} = \cdots = d^i_\beta = \Delta^u$), where $\Delta^a = \max\{|N_i^a||i \in V^a\}$ (resp., $\Delta^u = \max\{|N_i^u||i \in V^u\}$) is the maximum degree of G^a (resp., G^u).

• *Top-K reference distance:* For $i \in V^a$ (resp., V^u), its *Top-K reference distance feature* $\overline{f_K(i)}$ is a *K-dimensional vector* $(h_1^i, h_2^i, \cdots, h_K^i)$, where h_k^i $(1 \leq k \leq K)$ is the distance (the length of a shortest path) from i to the user with the k-th largest degree in G^a (resp., G^u). Note that it is possible $h_k^i = \infty$ if the graph is not connected.

• *Landmark reference distance:* Suppose $V_L^a = \{v_1, v_2, \cdots, v_L | v_k \in V^a\}$ is a set of users that has been de-anonymized (evidently, $V_L^a = \emptyset$ initially) to $U_L^u = \{u_1, u_2, \cdots, u_L | u_k \in V^u\}$ under some DA scheme σ with $\sigma(v_k) = u_k$ $(1 \leq k \leq L)$. Intuitively, V_L^a and U_L^u can be used as auxiliary information for future DA. Therefore, for $i \in V^a \setminus V_L^a$ (resp., $V^u \setminus U_L^u$), we define its *landmark reference distance feature* $\overline{f_l(i)} = (h_1^i, h_2^i, \cdots, h_L^i)$, where h_k^i $(1 \leq k \leq L)$ is the distance from i to $v_k \in V_L^a$ (resp., $u_k \in U_L^u$).

• *Sampling closeness centrality:* For $i \in V^a$ (resp., V^u), we define the *sampling closeness centrality feature* $f_c(i)$ to characterize its global topological property without inducing too much computational overhead. Formally, we first randomly sample a subset S^a of V^a (resp., S^u of V^u). Then, we define $f_c(i) = \sum_{j \in S^a \setminus \{i\}} \frac{1}{h(i,j)}$ (resp., $f_c(i) = \sum_{j \in S^u \setminus \{i\}} \frac{1}{h(i,j)}$), where $h(i,j)$ is the distance from i to j.

According to the aforementioned definitions, (i) we consider both local and global structural features of a user, e.g., the degree and neighborhood features characterize the local topological properties of a user while the Top-K reference distance and sampling closeness centrality features demonstrate the global topological characteristics of a user; (ii) we also consider the computational efficiency of obtaining these features for a user. For instance, instead of using the accurate *closeness centrality* of a user, we introduce a sampling closeness centrality feature, which can characterize the global feature of a user without causing too much computation overhead.

Now, based on the features defined for each user, we can quantitatively measure the *similarity* between an anonymized user $i \in V^a$ and a known user $j \in V^u$. Let $\overline{f_{d,c}(i)} = (f_d(i), f_c(i))$. Then, we define the *structural similarity* between $i \in V^a$ and

$j \in V^u$ as

$$\phi(i,j) = c_1 \cdot s(\overline{f_{d,c}(i)}, \overline{f_{d,c}(j)}) + c_2 \cdot s(\overline{f_n(i)}, \overline{f_n(j)}) \tag{84}$$

$$+ c_3 \cdot s(\overline{f_K(i)}, \overline{f_K(j)}) + c_4 \cdot s(\overline{f_l(i)}, \overline{f_l(j)}), \tag{85}$$

where $c_{1,2,3,4} \in [0,1]$ are constant values representing the weights and $c_1+c_2+c_3+c_4 = 1$, and $s(\cdot,\cdot)$ is the *Cosine similarity* between two vectors.

According to our theoretical quantification in Section 3.3, $\mathfrak{A}^*$ is inherently an optimization based algorithm with the objective of minimizing the DE Ψ_{σ_k}, which is different from most of existing DA algorithms (heuristics based) [27, 104, 127]. Inspired by our quantification, we design a novel and operational ***O**ptimization based **D**e-**A**nonymization* (**ODA**) scheme, which is a relaxed version of $\mathfrak{A}^*$.

In ODA, rather than using the DE function as in the quantification, we re-define $\psi_{i,j}$ and Ψ_σ as follows. Given a DA scheme $\sigma = \{(i,j)|i \in V^a, j \in V^u\}$, we define the DE on a user mapping $(i,j) \in \sigma$ as

$$\psi_{i,j} = |f_d(i) - f_d(j)| + (1 - \phi(i,j)) \cdot |f_d(i) - f_d(j)|, \tag{86}$$

and the DE on σ as

$$\Psi_\sigma = \sum_{(i,j)\in\sigma} \psi_{i,j}. \tag{87}$$

Based on Ψ_σ, we give the framework of ODA as shown in Algorithm 1. In Algorithm 1, $\Lambda^a \subseteq V^a$ is the target DA set and $\Lambda^u \subseteq V^u$ is the possible mapping set of Λ^a. GetTopDegree(X,y) is a function to return y users with the largest degree values in X, i.e., return $\{i|i$ has the Top-y degree in $X\}$. $\mathcal{C}(i) \subseteq \Lambda^u$ is the *candidate mapping set* for $i \in \Lambda^a$, which consists of the γ most possible mappings of i in Λ^u. GetTopSimilarity(i, Λ^u, γ) is a function to return γ users having the highest similarity scores ($\phi(i,\cdot)$) with i in Λ^u, i.e., return $\{j|j \in \Lambda^u,$ and j has the Top-γ $\phi(i,j)$ in $\Lambda^u\}$.

From Algorithm 1, ODA de-anonymizes G^a iteratively. During each iteration, ODA is trying to de-anonymize a subset of V^a and seeking the *sub-DA scheme* $\sigma^*(\Lambda^a)$

Algorithm 1: **O**ptimization based **D**e-**A**nonymization (**ODA**)

1 Define $\Lambda^a = \Lambda^u = \emptyset$;
2 **while** ***true*** **do**
3 $\quad \Lambda^a = \text{GetTopDegree}(V^a, \alpha)$, $\Lambda^u = \text{GetTopDegree}(V^u, \alpha)$;
4 $\quad$ for every $i \in \Lambda^a$, compute a *candidate mapping set* $\mathcal{C}(i) = \text{GetTopSimilarity}(i, \Lambda^u, \gamma)$;
5 $\quad$ apply the *consistent rule* and *pruning rule* to find the DA scheme $\sigma(\Lambda^a) \in \prod_{i \in \Lambda^a} (i \times \mathcal{C}(i))$ which induces the least DE $\Psi_{\sigma(\Lambda^a)}$, denoted by $\sigma^*(\Lambda^a) = \{(i_1, j_1), (i_2, j_2), \cdots, (i_\alpha, j_\alpha)\}$;
6 $\quad$ for each $(i, j) \in \sigma^*(\Lambda^a)$, **if** $\phi(i, j) \geq \theta$ **then**
7 $\quad\quad$ accept the mapping (i, j);
8 $\quad\quad$ $V^a = V^a \setminus \{i\}$, $V^u = V^u \setminus \{j\}$;
9 $\quad$ if no mapping in $\sigma^*(\Lambda^a)$ is accepted, ***break***;

which induces the least DE. We explain the idea of ODA in details as follows. In Line 3, we initialize the target DA set Λ^a and the candidate mapping set Λ^u. From the initialization, $|\Lambda^a|, |\Lambda^u| \leq \alpha$ (since it is possible $|V^a|, |V^u| \leq \alpha$), where α is an important parameter to control how many anonymized users will be processed each iteration. In Line 4, we compute a *candidate mapping set* $\mathcal{C}(i)$ for each $i \in \Lambda^a$. $\mathcal{C}(i)$ consists γ most similar users of i in Λ^u. Here, we define $\mathcal{C}(\cdot)$ mainly for reducing the computational complexity. Instead of trying every mapping from i to Λ^u, we only consider to map i to some user in $\mathcal{C}(i)$. Hence, γ is another important parameter to control the computational complexity of ODA. We will demonstrate how to set α and γ to make ODA computationally feasible in Theorem 9. In Line 5, we find a DA scheme $\sigma^*(\Lambda^a)$ on Λ^a such that $\Psi_{\sigma^*(\Lambda^a)} = \min\{\Psi_{\sigma(\Lambda^a)} | \sigma(\Lambda^a) \in \prod_{i \in \Lambda^a} (i \times \mathcal{C}(i))\}$, i.e., $\sigma^*(\Lambda^a)$ causes the least DE. Furthermore, the *consistent rule* and the *pruning rule* are applied in this step. The *consistent rule* makes any possible DA scheme $\sigma(\Lambda^a)$ consistent, i.e., no *mapping confliction* which is defined as the situation that two or more anonymized users are mapped to the same known user. This is because it is possible that $\mathcal{C}(i_1) \cap \mathcal{C}(i_2) \neq \emptyset$ for $i_1 \neq i_2 \in \Lambda^a$, and the situation $\sigma(i_1) = \sigma(i_2)$ in a DA scheme should be avoided. Note that, it possible that no $\sigma(\Lambda^a)$ is consistent.

In this case, we should increase γ to guarantee at least one $\sigma(\Lambda^a)$ is consistent. The *pruning rule* is used to remove some DA schemes whose DE is larger than the current known least DE. For instance, let $\sigma^*(\Lambda^a)$ be the DA scheme having the least DE after testing k possible DA schemes. Then, when testing the $(k+1)$-th possible DA scheme $\sigma_{k+1}(\Lambda^a)$, if partial of mappings in $\sigma_{k+1}(\Lambda^a)$ has already induced a larger DE than $\sigma^*(\Lambda^a)$, we stop test $\sigma_{k+1}(\Lambda^a)$ and continue the next one. On the other hand, if $\sigma_{k+1}(\Lambda^a)$ induces a smaller DE than $\sigma^*(\Lambda^a)$, we update $\sigma^*(\Lambda^a)$ to $\sigma_{k+1}(\Lambda^a)$. Both the consistent rule and the pruning rule can remove some unqualified DA schemes in advance, which can speed up ODA. Actually, although $\sigma^*(\Lambda^a)$ causes the least DE, $\sigma^*(\Lambda^a)$ is a local optimization solution (according to our quantification, the solution of $\mathfrak{A}^*$ is the optimum solution). This is because we try to seek a tradeoff between computational feasibility and DA accuracy. After obtaining $\sigma^*(\Lambda^a)$, we accept the mappings in $\sigma^*(\Lambda^a)$ with similarity scores no less than a *threshold value* θ (Lines 6-8). For the mappings that had been rejected, they will be re-considered in the following iterations for possible better DA. If no mapping can be accepted, we stop ODA. Subsequently, we analyze the time and space complexities of ODA in the following theorem.

Theorem 9. *(i) The space complexity of ODA is* $O(\min\{n^2, m+n\})$. *(ii) Let* γ *be some constant value,* $\alpha = \Theta(\log n)$, *and* Γ *be the average number of accepted mappings in each iteration of ODA. Then, the time complexity of ODA is* $O(m + n\log n + n^{\Theta(1)\log\gamma+1}/\Gamma)$ *in the worst case.*

Proof: (i) The space complexity of ODA is upper bound by $O(\min\{n^2, m+n\})$. The proof is straightforward and thus we omit it.

(ii) In ODA, we assume $f_d(i)$, $\overline{f_n(i)}$, $\overline{f_K(i)}$, $\overline{f_l(i)}$, and $f_c(i)$ are computed before the iteration starts. Then, the time consumption of computing these features is bounded by $O(m + n\log n)$. Then, from ODA, the worst case time consumption of each iteration is upper bounded by $\gamma^\alpha = \gamma^{\Theta(\log n)} = 2^{\log\gamma^{\Theta(\log n)}} = 2^{\Theta(\log n)\log\gamma} = n^{\Theta(1)\log\gamma}$.

Furthermore, the number of iterations in ODA is $\Theta(n/\Gamma)$. It follows the worst case time complexity of ODA is $O(m + n\log n) + O(n^{\Theta(1)\log\gamma+1}/\Gamma) = O(m + n\log n + n^{\Theta(1)\log\gamma+1}/\Gamma)$. □

Finally, we make some remarks on ODA as follows.

• ODA is a *cold start* algorithm, i.e., we do not need any priori knowledge, e.g., the seed mapping information [27, 104, 127], to bootstrap the DA process. Furthermore, unlike existing DA algorithms [27, 104, 127] which consist of two phases (*landmark/seed identification phase* and *DA propagation phase*), ODA is a single-phase algorithm. Interestingly, ODA itself can act as a *landmark identification algorithm*. From our experiment (Section 3.5.2), ODA can de-anonymize the 60-180 Top-degree users in Gowalla and Google+ (see Table 3) perfectly, which can serve as landmarks (V_L^a and U_L^u) for future DA. In addition, ODA as a landmark identification algorithm is much faster than that in [104] (with complexity of $O(nd^{k-1}) = O(n^k)$, where d is maximum degree of G^a/G^u and k is the number of landmarks) and [127] (with complexity of $k!$, could be computationally infeasible for a PC when $k \geq 20$).

• Similar to $\mathfrak{A}^*$, ODA is an optimization based DA scheme, which is different from most of existing heuristics based solutions [27, 104, 127]. In ODA, the objective is to minimize a DE function. The reasonableness and soundness of ODA lie on one direct conclusion of our theoretical quantification: *minimizing the DE leads to the best possible DA scheme*.

• In ODA, we seek an adjustable tradeoff between DA accuracy and computational feasibility. Although $\mathfrak{A}^*$ obtains the optimum solution *a.a.s.* in terms of our quantification, it is computationally infeasible ($O(n!)$). ODA has a polynomial time complexity of $O(m+n\log n+n^{\Theta(1)\log\gamma+1}/\Gamma)$ in the worst case, which is computationally feasible at the cost of sacrificing some accuracy. Based on our experiments on large scale real datasets in the following subsection, ODA is operable while preserves satisfiable DA performance.

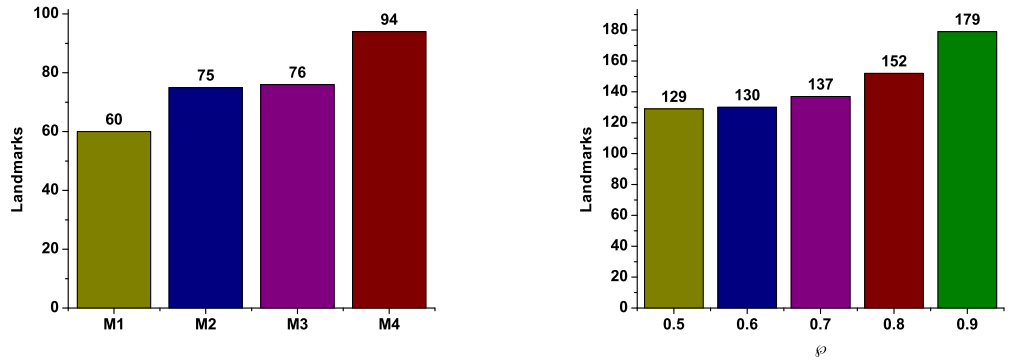
• ODA is a general framework. Line 5 can also be implemented by seeking a *maximum weighted bipartite graph matching* on a *weighted bipartite graph* $G(\Lambda^a \cup \Lambda^u, \bigcup_{i \in \Lambda^a} (i \times \mathcal{C}(i)))$, where the weight on each edge is $\phi(i, j)$ $(i \in \Lambda^a, j \in \mathcal{C}(i))$.

• In ODA, one implicit assumption is $V^a = V^u$, i.e., the G^a and G^u are defined on the same group of users. In practice, it is possible that V^a and V^u are not exactly the same. In this case, if V^a and V^u are not significantly different, ODA is also workable at the cost of some performance degradation ($(1-\epsilon)$-perfect DA). One better solution could be estimating the overlap between G^a and G^u first, and then apply ODA to the overlap to achieve better performance. We will take the estimation of the overlap between G^a and G^u as one of the future works.

3.5.2 Experimental Evaluation and Analysis

3.5.2.1 Datasets and Setup

We evaluate the performance of ODA on two real world datasets: Gowalla and Google+ (see the basic information in Section 3.4). Gowalla is a location based social network and consists of two different datasets [16, 115]. The first dataset is a spatiotemporal mobility trace consisting of 6,442,890 *check-ins* generated by 196,591 users. Each check-in has the format of <UserID, latitude, longitude, timestamp, location ID>. The second dataset is a social graph (950,327 edges) of the same 196,591 users. Assume the mobility trace is anonymized. Our objective is to de-anonymize the mobility trace using the social graph as auxiliary data. Since the mobility trace does not have an explicit graph structure, supposing the social graph is the ground truth, we apply the technique in [115] on the mobility trace to construct four graphs with different *recalls* and *precisions*, denoted by $M1, M2, M3$, and $M4$, respectively (recall $= \frac{\text{true positive}}{\text{true positive+false negative}}$ and precision $= \frac{\text{true positive}}{\text{true positive+false positive}}$). Particularly, the recall and precision of $M1$ are 0.6 and 0.865, of $M2$ are 0.72 and 0.83, of $M3$ are 0.75 and 0.78, and of $M4$ are 0.8 and 0.72, respectively. The second considering dataset is the Google+ dataset in Section 3.4, which has 4,692,671 users and



(a) Gowalla (b) Google+

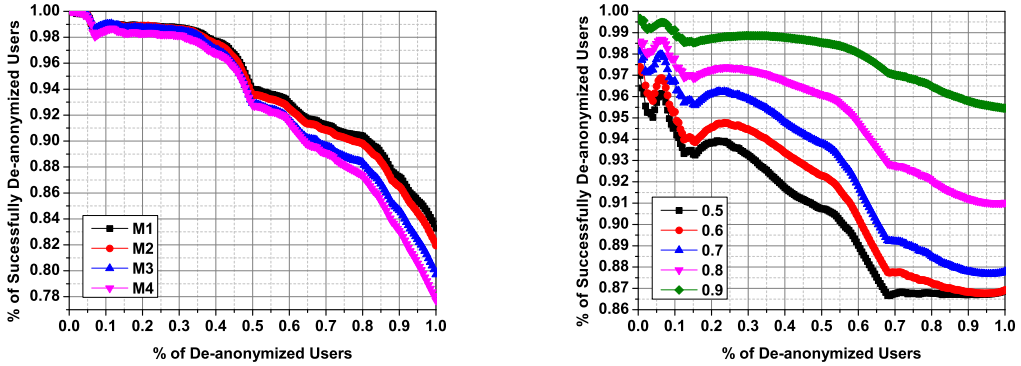
Figure 4: Landmark identification. $c_1, c_2 \in [0.1, 0.3], c_3 \in [0.4, 0.8], c_4 = 0, \alpha \in [10, 30], \gamma \in [1, 4]$.

90,751,480 edges. Given some projection probability $\wp \in [0.5, 0.9]$, We first use the *projection process* in Section 3.3 to produce G^a and G^u, and then use ODA to de-anonymize G^a with G^u as auxiliary data. Note that, the auxiliary data is from a different contextual domain (social data) with the anonymized data (mobility trace) in Gowalla while the auxiliary and anonymized data are from the same domain in Google+.

All the experiments are implemented on a PC with 64 bit Ubuntu 12.04 LTS operating system, Intel Xeon E5620 CPU (2.4GHz × 8 Threads), 48GB memory, and 2 disks with 8TB storage.

3.5.2.2 Results

Landmark Identification. As we mentioned in the previous subsection, ODA itself can work as a *landmark identification algorithm*. Let $V_L^a = U_L^u = \emptyset$ in ODA, i.e., $s(\overline{f_l(\cdot)}, \overline{f_l(\cdot)}) = 0$ in $\phi(\cdot, \cdot)$. Then, we run ODA on Gowalla and Google+ to identify some landmarks as shown in Fig. 4 (note that, the DA in ODA is conducted according to the degree non-increasing order). The results show that we can de-anonymize the first 60-94 users in Gowalla and the first 129-179 users in Google+ perfectly (100% correctly). For instance, when $G^a = M2$ in Gowalla, the first 75 users are perfectly



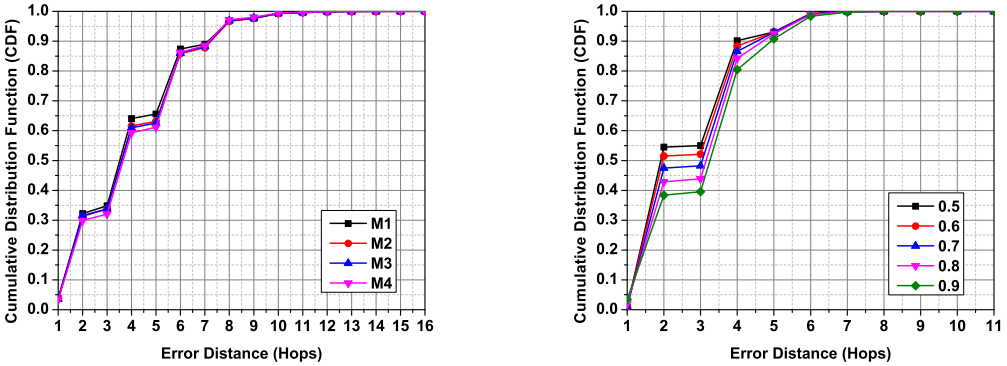
(a) De-anonymize Gowalla (b) De-anonymize Google+

Figure 5: De-anonymize Gowalla and Google+. $c_1, c_2 \in [0, 0.2], c_3 + c_4 \in [0.4, 1], \alpha \in [10, 30], \gamma \in [2, 10]$.

de-anonymizable and when $\wp = 0.7$, the first 137 users in Google+ are perfectly de-anonymizable. According to ODA, the identified landmarks can serve as references for future DA.

From Fig. 4 (a), we can see that when the recall increases, there are more common edges between G^a and G^u, which implies it is easier to identify the high degree users based on the increased structural information and thus more landmarks can be identified. Because of a similar reason, we can see from Fig. 4 (b) that more landmarks can be identified in Google+ for large $\wp$ due to more edge overlap between G^a and G^u.

De-anonymization Results. By taking the users identified in Fig. 4 as landmarks, we employ ODA to de-anonymize Gowalla $(M1, M2, M3, M4)$ and Google+ (G^a with different $\wp$) as shown in Fig. 5, where the x-axis represents the *accumulated percentage of users de-anonymized* and the y-axis represents the *accumulated percentage of users successfully de-anonymized*. From Fig. 5, we can see that the successful DA rate is higher for large-degree users than that of small-degree users, i.e., when x increases, the percentage of successfully deanonymized users generally show a decreasing trend. The reason is that large-degree users carry more structural



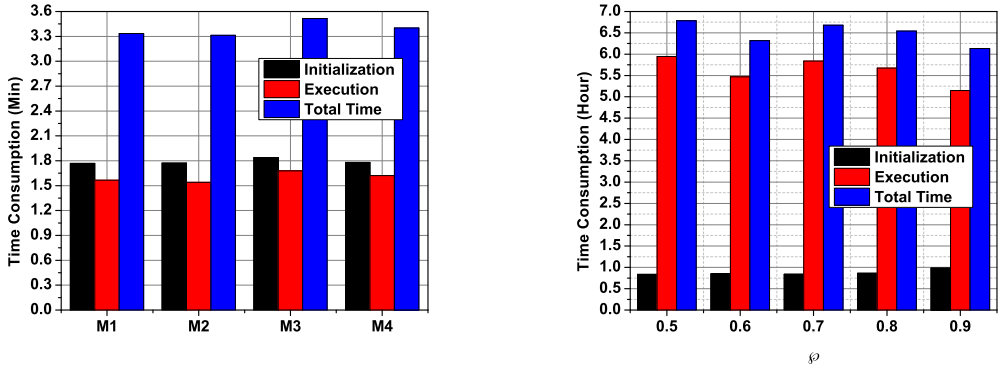
(a) Gowalla (b) Google+

Figure 6: DA error distribution.

information, which can thus be more accurately de-anonymizable. This can also be seen from our quantification. For Gowalla, we observe from Fig. 5(a) that although recall dominates the landmark identification process, the large-scale DA performance is impacted more by precision. Generally, a high precision implies this dataset is more de-anonymizable, e.g. $M4$. This is because a high precision implies a low false positive, which can be viewed as *noise* in practice, and thus the DA accuracy is better. For Google+, we see from Fig. 5 (b) that the G^a projected with a large $\wp$, e.g., $\wp = 0.9$, is more de-anonymizable. As shown in our quantification, this is because a large $\wp$ implies more similar between G^a and G^u and thus more users can be successfully de-anonymized.

From Fig. 5, we also see that the DA performance of ODA on Gowalla and Google+ is better than the evaluation results shown in Table 5, e.g., when $\wp = 0.9$, Table 5 indicates 91.2% of the users in Google+ are *a.a.s.* de-anonymizable while ODA successfully de-anonymizes 95.5% of the users. This is because the values shown in Table 5 are the lower bounds on de-anonymizable users. In summary, about $77.7\% - 83.3\%$ of the users in Gowalla and $86.9\% - 95.5\%$ of the users in Google+ are de-anonymizable in different scenarios. Thus, SDA is powerful in practice.

De-anonymization Error Analysis. For $i \in V^a$, let $i' \in V^u$ be i's correct



(a) Gowalla (b) Google+

Figure 7: Time consumption.

mapping. Suppose i is incorrectly mapped to some $j \in V^u$ (i.e., $j \neq i'$) by ODA. Then, the distance between i' and j in G^u is defined as the *error distance* corresponding to i. For Gowalla and Google+, we analyze the distribution of the error distance of all the incorrect mappings as shown in Fig. 6. From Fig. 6, we can see that almost all the incorrect mappings have an error distance of $2-10$ for Gowalla and of $2-6$ for Google+. Google+ has a smaller average error distance than Gowalla because its graph density is higher than Gowalla followed by shorter average distance among users. Furthermore, we can also see that the average error distance decreases with the increase of $\wp$ for Google+ in Fig. 6 (b). This is also because a large $\wp$ implies a dense graph followed by shorter average distance.

Time Consumption. We calculate the time consumption on de-anonymizing Gowalla and Google+ as shown in Fig. 7, where we provide the *initialization time* used for loading files and other initializations, the *execution time* used for executing the iterations in ODA, and *total time* consumed by ODA. Generally, the time consumption for Gowalla/Google+ is similar in different scenarios (the time difference may be caused by running other applications when run ODA). On average, the initialization time, execution time, and total time are 1.79 mins, 1.6 mins, and 3.39 mins for Gowalla and 0.88 hours, 5.61 hours, and 6.49 hours for Google+, respectively.

3.6 Implications and Discussion

Based on our DA quantification, evaluation results on 26 real world datasets, and DA practice ODA, we provide some implications in this section. We also discuss the impacts of our findings to *secure data publishing* in practice and the guidelines for future data publishing.

Structural information may induce privacy leakage. Although we have some practices (including ODA) that show SDA is possible, in this chapter, we theoretically demonstrate the reasons by providing rigorous quantification under a general data model. From the quantification, structural information can enable large-scale perfect or $(1-\epsilon)$-perfect DA. Therefore, *for secure data publication, besides the data itself, the information carried by data structure is also essential and deserves dedicated consideration and efforts.*

The fact is that we still have a long way to go to achieve secure data publishing. From our large scale study on 26 real world datasets, most of existing graph datasets are de-anonymizable based only on their structural information. On the other hand, existing anonymization techniques are vulnerable to SDA attacks. Therefore, *new anonymization techniques should be developed.* Meanwhile, since graph data release/sharing/trasferring has significant business and social value, *the data utility should be preserved in the new developed anonymization schemes.* In summary, we are expected to develop *new secure data publishing schemes that properly achieve a balance between data privacy protection and data utility preservation.*

Suggestions for secure data publishing. Secure data publishing is important for businesses, research, and the society. However, with the wide availability of richer auxiliary information, especially with the emerging of *Collaborative Information Seeking* (CIS) systems and *data/knowledge brokers* [124, 148], the privacy of people, businesses, governments, etc. will increasingly be compromised. For secure data publishing, we have some general suggestions as follows. (i) *Carefully sharing*

data with or transferring data to third parties and partners. Before sharing the data, the data owners should examine the dedicated applications to see if the data sharing is necessary. Based on the requirements of applications, the data could be shared in different granularity levels: *digest level*: share/transfer a digest/summary of the data to third parties or partners; *partial and density-control level*: based on our quantification, controlling the graph density could increase the difficulty of DA. Therefore, in this level, only a density-controlled anonymized version (e.g., by sampling) of a subset of the data (e.g., a community) is shared/transferred; *density-control level*: a density-controlled anonymized version of the data is shared/transferred; *full level*: an anonymized version of the full dataset is shared/transferred. (*ii*) *Evaluate the potentially vulnerability of the dataset before actual publishing.* Before actually publish the data, the data owners can evaluate the vulnerability of the data. For instance, if the data is graph data, the data can be evaluated using our quantification as in Section 3.4. (*iii*) *Develop proper policy on data collection.* Many graph data owners allow public data collection, e.g., Twitter, Facebook. allow crawlers and other automatic programs to collect users information online. This could increase the data DA risk by providing auxiliary information to adversaries. Therefore, it is better for data owners to develop proper policies to limit such public data collection.

3.7 Chapter Summary

In this chapter, we study the quantification, practice, and implications of graph data DA. First, for the first time, we address several fundamental open problems in the data DA research by quantifying the conditions for *perfect DA and* $(1-\epsilon)$*-perfect DA under a general data model.* This remedies the gap between graph data DA practice and theory. Second, we conduct a large scale study on the de-anonymizability of 26 diverse real world graph datasets, which turn out to be de-anonymizable partially or perfectly. We also quantitatively demonstrate the necessary conditions and reasons

for the de-anonymizability of the 26 datasets. Third, following our quantification, we propose a practical DA technique that is a *cold start single-phase Optimization based De-Anonymization* (ODA) algorithm. We also analyze ODA theoretically and experimentally. The experimental results show that $77.7\% - 83.3\%$ of the users in Gowalla (196,591 users, 950, 327 edges) and $86.9\% - 95.5\%$ of the users in Google+ (4,692,671 users, 90,751,480 edges) can be de-anonymized, which implies SDA is implementable and powerful in practice. Finally, we conclude some implications from our findings.

CHAPTER IV

SEED-BASED DE-ANONYMIZATION QUANTIFICATION

4.1 Introduction

Due to the vulnerability of existing anonymization schemes, the emerging *Structure based De-Anonymization* (SDA) attacks have been experimentally demonstrated to break the privacy of graph data effectively only based on the data's structural information, e.g., Narayanan and Shmatikov's De-Anonymization (DA) attack [104], Srivatsa and Hicks' DA attack [127]. Furthermore, there is some preliminary analysis on the de-anonymizability of graph data under the the *Erdös-Rényi* (ER) random graph model or the *preferential attachment* model [69, 113, 151]. On one hand, these existing analyses shed light on the research of quantifying the de-anonymizability of graph data. On the other hand, however, all the existing analyses have some limitations, e.g., some did not consider the seed information, the use of an unrealistic network model, unrealistic assumptions, overlooked other more powerful structural information. These limitations prevent most existing analyses to be applicable to real world graph data. Aiming at addressing the limitations of existing de-anonymizability quantification techniques, we study the seed-based de-anonymizability of graph data in this chapter. Specifically, our contributions can be summarized as follows.

1. To the best of our knowledge, we conduct the first seed-based theoretical quantification on the *perfect de-anonymizability* and *partial de-anonymizability* of graph data under the ER model as well as in general scenarios, where the graph can

[0]Without of specification, "de-anonymization" means "seed-based de-anonymization" in this chapter.

follow an arbitrary data model. Therefore, our quantification can be applied to real world graph data and can quantitatively demonstrate the vulnerability of real world graph data to existing seed-based DA attacks. More importantly, our quantification provides the theoretical foundation for existing seed-based DA attacks, which closes the gap between practice and theory.

2. Based on our quantification, we implement a large scale evaluation on the perfect and partial de-anonymizability of 24 various real world social networks. In our evaluation, we show the conditions on perfectly and partially de-anonymizing a social network; how de-anonymizable a social network is according to its topological properties; and how many users of a social network can be successfully de-anonymized. Our evaluation results demonstrate that most of social networks, if not all, can be perfectly or at least partially de-anonymized depending on their structural properties.

3. Based on our quantification and evaluation, we find that compared to the structural information associated with known seed users, the other structural information (the structure among anonymized users) is more useful in improving DA attacks. We show that, both theoretically and experimentally, the overall structural information based DA is more powerful than seed-based DA, and a graph dataset is perfectly or partially de-anonymizable even without any seed information. As a result, this finding provides the foundation of an implication that one can design new effective DA attacks without seed information.

The rest of this chapter is organized as follows. In Section 4.2, we describe the system model, assumptions, and problem definition. The preliminary quantification under the ER model is implemented in Section 4.3. We conduct the quantification on perfect and partial seed-based de-anonymizability of graph data in general scenarios in Section 4.4. In Section 4.5, we evaluate the de-anonymizability of 24 real world social networks. Finally, the chapter is concluded in Section 4.6.

Table 8: Summarization of notations.

notation	definition
$G^a = (V^a, E^a)$	anonymized graph
$G^u = (V^u, E^u)$	auxiliary graph
i, j	nodes/users
n	number of users
$e^a_{i,j}, e^u_{i,j}$	user tie (links/edges)
d^a_i, d^u_i	degree of i
s_a, s_u, s	graph sampling probabilities
σ	a DA scheme
σ_0	the perfect DA
σ_k	a DA scheme with k errors
$\mathcal{S}$	seed mappings
$\Lambda = \lvert\mathcal{S}\rvert$	the cardinality of $\mathcal{S}$
$\Delta_{\sigma:(i,j)}$	edge difference induced by $(i,j) \in \sigma$
Δ_σ	edge difference of σ
$G(n,p)$	ER random graph with parameters n and s
ϵ	tolerated DA error
m	number of edges
ρ, ρ_U	graph density
$\gamma_{U,W}$	graph connectivity

4.2 System Model, Assumption, and Definition

In this section, we introduce the system model, assumptions, and definitions. Generally, we employ similar data model, assumptions, and definitions as in Chapter 3. For conveniently reference and for the purpose of the discussion in this chapter, we restate the data model, assumptions, and definitions. To improve the readability, we summarize the frequently used acronyms and symbols in Table 8.

Data Model. In our quantification and evaluation, we employ the same graph model as in [27, 63, 64, 69, 104] to represent graphs. Specifically, the anonymized data is modeled by graph $G^a = (V^a, E^a)$, where $V^a = \{i | i \text{ is an anonymized user}\}$ and $E^a = \{e^a_{i,j} | i, j \in V^a, \text{ a tie exists between } i \text{ and } j\}$. To de-anonymize G^a, we use an auxiliary graph which has overlap users with G^a and can be obtained through

multiple manners, e.g., data aggregation, data mining, collaborative information systems, knowledge/data brokers [27, 63, 69, 104, 124, 148][1]. The auxiliary data is also modeled by a graph $G^u = (V^u, E^u)$, where $V^u = \{i | i \text{ is a known user}\}$ and $E^u = \{e^u_{i,j} | i, j \in V^u, \text{a tie exists between } i \text{ and } j\}$. To conduct the theoretical quantification without involving too much mathematical details, we assume both G^a and G^u are undirected graphs[2]. Furthermore, since our quantification and evaluation are based on the graph model, our work can be potentially applied to other kinds of data which can be modeled by graphs.

Given $i \in V^a$, its *neighborhood* is defined as $N^a_i = \{j | j \in V^a \wedge \exists e^a_{i,j} \in E^a\}$. Then, we define $d^a_i = |N^a_i|$ as the *degree* of i. Similarly, for $j \in V^u$, we can define its *neighborhood* N^u_j and *degree* d^u_j.

Graph Sampling. To make the quantification mathematically tractable, we employ the same assumptions on G^a and G^u in [69, 113, 151]. First, $V^a = V^u = \{1, 2, \cdots, n\}$ [69, 113, 151]. In the case that $V^a \neq V^u$, we can simply satisfy this assumption by adding the users in $V^u \setminus V^a$ to V^a and adding the users in $V^a \setminus V^u$ to V^u without changing E^a or E^u, i.e. adding the dissimilar users to each other with degree zero to make V^a and V^u are mathematically equivalent. Note that this is only a mathematical assumption without limiting the generality of this work. Our quantification is also valid in the case $V^a \neq V^u$.

[1]For the detailed means of obtaining the auxiliary data, please refer to the discussion in [63, 104]. Especially, with the emergence of data brokers, many auxiliary data can be easily obtained with an affordable cost.

[2]In reality, many graph data carry direction information, i.e., they are directed graphs. Furthermore, some DA attacks are designed to utilize the direction information to improve the DA performance, e.g., [104]. In this chapter, we do not take into account the direction information. The main reason is that we want to make our quantification sufficiently general. Although our quantifications are based on the undirected graph model, they can be extended to directed graphs directly by overlooking the direction information on edges.

Nevertheless, when applying our quantifications to directed graphs, the overlooking of the direction information may lead to inaccurate de-anonymizability quantification (potentially underestimate the de-anonymizability of the data). The problem of studying the impact of direction information on the de-anonymizability of graph data itself is an interesting research topic, which requires a proper model to characterize the direction information, elegant quantification techniques, and dedicated research. We take this research as one of our future research directions.

Second, based on the first assumption, we assume that G^a *and* G^u *are two sampling versions of an underlying conceptual graph* $G = (V, E)$ *in the physical world, where* $V = V^a = V^u$ *and* E *is the set of the true relationships among users in* V [69, 113, 151]. Particularly, we assume G^a is sampled from G by *independently and identically* sampling each edge in E with probability s_a, i.e., for $\forall e_{i,j} \in E$, $\Pr(e_{i,j} \in E^a | e_{i,j} \in E)$. Similarly, G^u is another sampled version of G with probability s_u. This assumption is also reasonable since people usually involve in multiple computer system contexts (e.g., social networks) and G^a and G^u are some particular graphs of users in V. For instance, G^a could be LinkedIn (a professional social network of V) while G^u is Facebook (a friendship social network of V).

De-anonymization. Based on our data model, a DA scheme can be formally defined as a mapping: $\sigma : G^a \to G^u$. Under σ, $\forall i \in V^a$, its mapping is $\sigma(i) \in V^u$. Since $V^a = V^u$, for simplicity, we define a *successful DA* of $i \in V^a$ is achieved under σ if $i = \sigma(i)$. In addition, we use σ_0 to denote the *perfect DA*, i.e., $\sigma_0 = \{(i, i) | i = 1, 2, \cdots, n\}$ (all the users of in G^a are correctly de-anonymized), and σ_k to denote any DA scheme with k incorrect mappings, i.e., k users are incorrectly de-anonymized under σ_k. Evidently, $k \in [2, n]$. In the rest of this chapter, we say that $i \in V^a$ is perfectly de-anonymizable if i can be correctly de-anonymized and V^a is perfectly de-anonymizable if all the users in V^a can be correctly de-anonymized.

Most existing DA algorithms (e.g., [27, 104, 127]) consist of two phases: *seed identification phase* which identifies some *seed mapping information* from V^a to V^u and *mapping propagation phase* which propagates the seed mapping information to de-anonymize the rest of the anonymized users. In this chapter, we focus on quantifying the de-anonymizability of graph data with seed knowledge. Therefore, as in [27, 104, 127], we assume we have identified a *seed mapping set* from V^a to V^u by some technique (e.g., the methods in [27, 104, 127]), denoted by $\mathcal{S} = \{(i, \sigma(i)) | i \in V^a, \sigma(i) \in V^u, i = \sigma(i)\}$. Furthermore, we define $\Lambda = |\mathcal{S}|$ as the number of seed mappings. For

convenience, we denote the seed users in V^a and V^u as $\mathcal{S}^a = \{i|(i,\sigma(i)) \in \mathcal{S}\}$ and $\mathcal{S}^u = \{i|(\sigma^{-1}(i),i) \in \mathcal{S}\}$, respectively. Then, our problem now is to quantify the de-anonymizability of a graph G^a given $\mathcal{S}$, G^u, and the existing of G, s_a, and s_u.

To make the quantification easy to follow and the conclusions succinct, we further assume $s_a = s_u = s$, i.e., we assume G^a and G^u are two instances of G with the same sampling probability. Note that, this assumption does not change our analysis in any material detail. All our quantification results can be extended to the case $s_a \neq s_u$ only with more complex expressions.

Measuring σ. Given G^a, G^u, and a DA scheme σ, we measure σ by the *edge difference* between G^a and G^u under σ. First, $\forall e^a_{i,j} \in E^a$, we define $\sigma(e^a_{i,j}) = e^u_{\sigma(i),\sigma(j)}$. Furthermore, let $E^a_i(A \subseteq V^a) = \{e^a_{i,v}|v \in N^a_i \cap A\}$, and $\sigma(E^a_i(A)) = \{\sigma(e^a_{i,v})|e^a_{i,v} \in E^a_i(A)\}$ ($\sigma(e^u_{i,j})$, $E^u_i(A)$, and $\sigma^{-1}(E^u_i(A))$ are defined in the same way). Specifically, let $E^a_i = E^a_i(V^a)$ and $E^u_j = E^u_j(V^u)$ for convenience. Then, we can define the edge difference induced by mapping $(i,\sigma(i) = j) \in \sigma$ as

$$\Delta_{\sigma:(i,j)} = |\sigma(E^a_i) \setminus E^u_j| + |\sigma^{-1}(E^u_j) \setminus E^a_i|, \tag{88}$$

i.e., $\Delta_{\sigma:(i,j)}$ measures the edge difference of users i and j under σ. Based on $\Delta_{\sigma:(i,j)}$, we measure σ by

$$\Delta_\sigma = \sum_{(i,j)\in\sigma} \Delta_{\sigma:(i,j)}, \tag{89}$$

which indicates the edge difference between G^a and G^u under σ. Intuitively, since G^a and G^u are strongly correlated (highly similar), it is expected that $\Delta_{\sigma_0} \leq \Delta_{\sigma_k}$ for $k \in [2,n]$ (we demonstrate this conclusion in Sections 4.3 and 4.4).

Similar as $\Delta_{\sigma:(i,j)}$ and Δ_σ, we define $\Delta_{\sigma:(i,j)}(\mathcal{S})$ which measures the the edge difference of a mapping (i,j) with respect to $\mathcal{S}$:

$$\Delta_{\sigma:(i,j)}(\mathcal{S}) = |\sigma(E^a_i(\mathcal{S}^a) \setminus E^u_j(\mathcal{S}^u)| + |\sigma^{-1}(E^u_j(\mathcal{S}^u) \setminus E^a_i(\mathcal{S}^a)|,$$

and $\Delta_\sigma(\mathcal{S})$ which measures the edge difference of a DA scheme σ with with respect to $\mathcal{S}$:

$$\Delta_\sigma(\mathcal{S}) = \sum_{(i,j)\in\sigma} \Delta_{\sigma:(i,j)}(\mathcal{S}). \tag{90}$$

4.3 Quantification under the Erdös-Rényi Model

In this section, we quantify the de-anonymizability of G^a with $\mathcal{S}$, G^u, G, and s under the Erdös-Rényi (ER) model, i.e. we assume $G(V, E)$ is a *random graph* generated from the ER model $G(n, p)$, where n is the number of nodes in the graph and p specifies the probability of an edge existing between any pair of nodes. Although real world graph data rarely satisfy the ER model [107], the analysis in this section can shed the light of the quantification in general scenarios (Section 4.4).

4.3.1 $\mathcal{S}$ based Quantification

As a warm up, we first quantify the de-anonymizability of G^a only based on the seed information $\mathcal{S}$. For the DA scheme σ, we assume σ *de-anonymizes each user* $i \in V^a \backslash \mathcal{S}^a$ *to some user* $\sigma(i) \in V^u \backslash \mathcal{S}^u$ such that $(i, \sigma(i))$ induces the least $\Delta_{\sigma:(i,\sigma(i))}(\mathcal{S})$ [3].

Theorem 10. *If* $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{2\ln n+1}{\Lambda}$ *(i.e.,* $\Lambda \geq \frac{4(2\ln n+1)(2-s-ps)}{ps^3(1-p)^2}$*), then it is* asymptotically almost surely (a.a.s.)[4] *that* $\forall i \in V^a \setminus \mathcal{S}^a$*, i is perfectly de-anonymizable (i can be successfully de-anonymized).*

Proof: For $\forall i \in V^a \setminus \mathcal{S}^a$ and $\forall v \in \mathcal{S}^a$, $\Pr(e^a_{i,v} \in E^a) = ps$. This is because $\Pr(e_{i,v} \in E) = p$ ($G(V, E)$ is a ER random graph $G(n, p)$) and $\Pr(e_{i,v}$ is sampled into $E^a | e_{i,v} \in E) = s$. Similarly, $\Pr(e^u_{\sigma(i),v} \in E^a) = ps$. Now, given a DA scheme

[3]Since our focus is on quantifying the de-anonymizability of G^a, we do not consider the actual DA algorithms. Specifically, we are aiming at providing the theoretical foundation on the workability of seed-base SDA attacks, e.g., [27, 104, 127].

[4]Asymptotically almost surely (a.a.s.) *implies that* an event happens with probability goes to 1 as $n \to \infty$.

σ, if $(i, \sigma(i)) = (i, i)$, i.e., i is correctly de-anonymized under σ, then, for a possible edge $e_{i,v} \in E$, it induces an edge difference if $e_{i,v}$ exists and $e_{i,v}$ is sampled into exactly one of E^a and E^u, which has a probability of $2ps(1-s)$; on the other hand, if $(i, \sigma(i)) = (i, j \neq i)$, i.e., i is incorrectly de-anonymized under σ, then, for an possible edge $e_{i,v} \in E$, it induces an edge difference with probability $2ps(1-ps)$. Let $\Delta_{\sigma:(i,\sigma(i))}(\mathcal{S}) = X$ if $\sigma(i) \neq i$ and $\Delta_{\sigma:(i,\sigma(i))}(\mathcal{S}) = Y$ if $\sigma(i) = i$. Then, we have

$$X \sim \mathbf{B}(\Lambda, 2ps(1-ps)), \text{if } \sigma(i) \neq i \tag{91}$$

$$Y \sim \mathbf{B}(\Lambda, 2ps(1-s)), \text{if } \sigma(i) = i \tag{92}$$

where $\mathbf{B}(x, y)$ is a *binomial distribution* with parameters x and y. Then, the mean values of X and Y are $\lambda_X = 2ps(1-ps)\Lambda$ and $\lambda_Y = 2ps(1-s)\Lambda$, respectively. Applying Lemma 1, we have

$$\Pr(X - Y \leq 0) \leq 2\exp(-\frac{1}{8}\frac{[2ps(1-ps)\Lambda - 2ps(1-s)\Lambda]^2}{2ps(1-ps)\Lambda + 2ps(1-s)\Lambda}) \tag{93}$$

$$= 2\exp(-\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps}\Lambda) \tag{94}$$

$$\leq \exp(-2\ln n - 1) \tag{95}$$

$$< \frac{1}{n^2}. \tag{96}$$

Since $\sum\limits_{n>0} \frac{1}{n^2} = \frac{\pi^2}{6} < \infty$, it is a.a.s. that $\Pr(X - Y \leq 0) \underset{n \to \infty}{\to} 0$, i.e., with probability goes to 1, the correct DA of i leads to the least $\Delta_{\sigma:(i,\sigma(i))}(\mathcal{S})$. Since the number of possible mappings $(i, \sigma(i))$ is upper bounded by $|V^u \setminus \mathcal{S}^u|$, it a.a.s. that i is perfectly de-anonymizable. □

In Theorem 10, we quantify the condition on p, s, and $\mathcal{S}$ on perfectly de-anonymizing any user in $V^a \setminus \mathcal{S}^a$. Now, we quantify the condition requirement for a stronger conclusion in Theorem 11, which indicates the condition on p, s, and $\mathcal{S}$ such that all the users in $V^a \setminus \mathcal{S}^a$ are perfectly de-anonymizable.

Theorem 11. *If* $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{2\ln n + \ln(2(n-\Lambda))}{\Lambda}$ *(i.e.,* $\Lambda \geq \frac{4(2\ln n + \ln(2(n-\Lambda)))(2-s-ps)}{ps^3(1-p)^2}$*), it is a.a.s. that all the users in* $V^a \setminus \mathcal{S}^a$ *are perfectly de-anonymizable.*

Proof: To prove this theorem, it is sufficient to prove that $\exists\sigma$ *such that* σ *perfectly de-anonymizes all the users in* $V^a \setminus \mathcal{S}^a$ *in bounded time.* As in Theorem 10, in σ, we de-anonymize $i \in V^a \setminus \mathcal{S}^a$ to $\sigma(i) \in V^u \setminus \mathcal{S}^u$ such that $(i, \sigma(i))$ induces the least $\Delta_{\sigma:(i,\sigma(i))}(\mathcal{S})$. Let $\mathbf{E}$ be the event that *there is at least one incorrectly de-anonymized user in* $V^a \setminus \mathcal{S}^a$, and X and Y are as defined in the proof of Theorem 10. Then, based on *Boole's inequality*, we have

$$\Pr(\mathbf{E}) \leq \sum_{i \in V^a \setminus \mathcal{S}^a} \Pr(i \text{ is incorrectly de-anonymized}) \tag{97}$$

$$= \sum_{i \in V^a \setminus \mathcal{S}^a} \Pr(X \leq Y) \tag{98}$$

$$\leq \sum_{i \in V^a \setminus \mathcal{S}^a} 2\exp(-\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps}\Lambda) \tag{99}$$

$$= 2(n-\Lambda)\exp(-\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps}\Lambda) \tag{100}$$

$$\leq 2(n-\Lambda)\exp(-2\ln n - \ln(2(n-\Lambda))) \tag{101}$$

$$= \frac{1}{n^2}. \tag{102}$$

Therefore, it is a.a.s. that $\Pr(\mathbf{E}) \underset{n \to \infty}{\to} 0$, with probability 1, all the users in $V^a \setminus \mathcal{S}^a$ are perfectly de-anonymizable when $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{2\ln n + \ln(2(n-\Lambda))}{\Lambda}$. □

4.3.2 Sophisticated Quantification: Considering more Structure Information

In the previous subsection, we quantified the de-anonymizability of G^a based only on the seed knowledge. Actually, besides the edges in $E_i^a(\mathcal{S})/E_i^u(\mathcal{S})$, all the edges in E_i^a/E_i^u can provide structure information which can be used for DA. In this subsection, we consider to quantify the de-anonymizability of G^a based on all the adjacent edges of $i \in V^a$, i.e., we consider both the structural information carried by seed mappings in $\mathcal{S}$ and the overall topological information of G^a and G^u. First, we quantify the structural conditions on G^a and G^u for perfect DA in Theorem 12. Theorem 12 has two parts. The first part shows the condition such that $\Delta_{\sigma_0} < \Delta_{\sigma_k}$ for any given

σ_k. The second part demonstrates the condition for a much stronger conclusion such that σ_0 is the one and the only one inducing the least edge difference. Basically, the first part of Theorem 12 can be proven using a similar technique as in [113]. Here, we obtain a tighter bound by applying more elegant quantification techniques.

Theorem 12. *(i) If* $\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{2\ln n+1}{k(n-k/2-1)}$, *it is a.a.s. that* $\Delta_{\sigma_0} < \Delta_{\sigma_k}$ $(k \in [2,n])$, *i.e., it is a.a.s. that the perfect DA scheme* σ_0 *induces less edge difference than any given DA scheme* $\sigma_k \neq \sigma_0$*; (ii) If* $\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{(k+2)\ln n+\ln(2(n-\Lambda-1))}{k(n-k/2-1)}$, *it is a.a.s. that the perfect DA scheme* σ_0 *induces the least edge difference than all the other DA schemes, i.e., it is a.a.s. that* σ_0 *is the only scheme inducing the least edge difference.*

Proof: (i) First, let $V_k \subseteq V$ be the set of users that are incorrectly de-anonymized under σ_k, $V_0 = V \setminus V_k$ be the set of users been correctly de-anonymized, $E_k = \{e_{i,j} | i \in V_k \vee j \in V_k\}$ be the set of all possible edges that adjacent to at least one incorrectly de-anonymized users, and $E_0 = \{e_{i,j} | i,j \in V_0\}$ be the set of all possible edges among the correctly de-anonymized users. Furthermore, let $m_k = |E_k|$ and $m_0 = E_0$. Therefore, we have $m_k = \binom{k}{2} + k(n-k)$ and $m_0 = \binom{n-k}{2}$. Now, we define two random variables $X = \Delta_{\sigma_k}$ and $Y = \Delta_{\sigma_0}$.

It is evident that if an edge is sampled into exactly one of G^a and G^u, this edge will induce one edge difference in Δ_{σ_0}. Consequently, Y is a *binomial variable* with parameters $m_0 + m_k$ and $2ps(1-s)$, i.e.

$$Y \sim \mathbf{B}(m_0 + m_k, 2ps(1-s)). \tag{103}$$

Similarly, under σ_k, each edge in E_0 will induce an edge difference if it is sampled into exactly one of G^a and G^u. For each edge in E_k, if it is not a *transposition edge*[5], it will cause an edge difference with probability of $2ps(1-ps)$; otherwise, if is a transposition edge, it will cause an edge difference with probability of $2ps(1-s)$. Since we have at

[5]A *transposition edge* is an edge such that the two endpoints of this edge are incorrectly de-anonymized to each other, i.e., a *transposition edge* an edge $e_{i,j}$ such that $(i,j) \in \sigma_k \wedge (j,i) \in \sigma_k$.

most $k/2$ transposition edges, we have

$$X \underset{\text{stochastically}}{\geq} \mathbf{B}(m_0, 2ps(1-s)) + \mathbf{B}(m_k - k/2, 2ps(1-ps)) + \mathbf{B}(k/2, 2ps(1-s)). \tag{104}$$

Let $\widetilde{X} \sim \mathbf{B}(m_k - k/2, 2ps(1-ps))$ and $\widetilde{Y} \sim \mathbf{B}(m_k - k/2, 2ps(1-s))$ be two binomial random variables. Then, we have

$$\Pr(X \leq Y) \underset{\text{stochastically}}{=} \Pr(\widetilde{X} \leq \widetilde{Y}). \tag{105}$$

Applying Lemma 1, we have

$$\begin{aligned}
\Pr(X \leq Y) &\underset{\text{stochastically}}{=} \Pr(\widetilde{X} \leq \widetilde{Y}) && (106)\\
&\leq 2\exp(-\frac{(\lambda_{\widetilde{X}} - \lambda_{\widetilde{Y}})^2}{8(\lambda_{\widetilde{X}} + \lambda_{\widetilde{Y}})}) && (107)\\
&= 2\exp(-\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps}(m_k - k/2)) && (108)\\
&\leq 2\exp(-2\ln n - 1) && (109)\\
&< \frac{1}{n^2}. && (110)
\end{aligned}$$

Since $\sum_{n>0} \frac{1}{n^2} = \frac{\pi^2}{6} < \infty$, it is a.a.s. that $X > Y$, i.e., $\Delta_{\sigma_0} < \Delta_{\sigma_k}$, for any $\sigma_k \neq \sigma_0$.

(ii) Let $\mathbf{E}$ be the event that *there is some DA scheme* σ_k *such that* $\sigma_k \neq \sigma_0$ *and* $\Delta_{\sigma_k} \leq \Delta_{\sigma_0}$. Then, applying the union bound on $\Pr(\mathbf{E})$, we have

$$\begin{aligned}
\Pr(\mathbf{E}) &= \bigcup_{k=2}^{n-\Lambda} \Pr(\Delta_{\sigma_k} \leq \Delta_{\sigma_0}) && (111)\\
&\leq \sum_{k=2}^{n-\Lambda} n^k \cdot \Pr(\Delta_{\sigma_k} \leq \Delta_{\sigma_0}) && (112)\\
&\leq \sum_{k=2}^{n-\Lambda} n^k \cdot 2\exp(-\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps}(m_k - k/2)) && (113)\\
&\leq \sum_{k=2}^{n-\Lambda} n^k \cdot 2\exp(-(k+2)\ln n - \ln(2(n-\Lambda-1))) && (114)\\
&= \frac{1}{n^2}. && (115)
\end{aligned}$$

Consequently, it is a.a.s. that $\Pr(\mathbf{E}) \sim 0$ as $n \to \infty$, i.e., it is a.a.s. that σ_0 is the only scheme that induces the least edge difference. □

Theorem 12 has a very strong implication: *even without any seed information, it still possible to perfectly de-anonymize a large scale graph.* We summarize this implication in Corollary 1.

Corollary 1. *If* $\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{(k+2)\ln n+\ln(2(n-1))}{k(n-k/2-1)}$, *it is a.a.s. that the perfect DA scheme* σ_0 *induces the least edge difference than all the other DA schemes, i.e., it is a.a.s. that* σ_0 *is the only scheme inducing the least edge difference.*

Based on Theorems 11, 12 and Corollary 1, it is straightforward to obtain a more accurate (tighter) bound on the structure condition of G^a and G^u for perfect DA as shown in Theorem 13.

Theorem 13. *If* $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \min\{\frac{2\ln n+\ln(2(n-\Lambda))}{\Lambda}, \frac{(k+2)\ln n+\ln(2(n-\Lambda-1))}{k(n-k/2-1)}\}$, *where* $\Lambda \in [0, n]$, G^a *is perfectly de-anonymizable.*

4.3.3 Quantification with Error Toleration

Now, we study the structural condition on G^a and G^u given $\mathcal{S}$ such that some *DA error* is tolerated. Let $\epsilon \in [0, 1-\frac{\Lambda}{n}]$ be some constant value. We define G^a is $(1-\epsilon)$-*de-anonymizable* if at least $(1-\epsilon)n$ users in G^a are perfectly de-anonymizable. Then, we specify the condition such that G^a is $(1-\epsilon)$-deanonymizable with or without seed information in Theorem 14, i.e., the condition that at most ϵn incorrect DA are allowable.

Theorem 14. *If* $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \min\{\frac{2\ln n+\ln(2(n-\epsilon n-\Lambda))}{\Lambda}, \frac{(k+2)\ln n+\ln(2(n-\epsilon n-\Lambda))}{k(n-k/2-1)}\}$, *where* $\Lambda \in [0, n]$, *then* G^a *is* $(1-\epsilon)$*-de-anonymizable.*

Proof: First, we prove that if $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{2\ln n+\ln(2(n-\epsilon n-\Lambda))}{\Lambda}$, G^a is $(1-\epsilon)$-de-anonymizable. Let $V_c \subseteq V^a \setminus \mathcal{S}^a$ and $|V_c| = n - \epsilon n - \Lambda$. Furthermore, let $\mathbf{E}$ be the

event that *there is at least one incorrectly de-anonymized user in* V_c. Then, using the similar proof technique in Theorem 11, we have

$$\Pr(\mathbf{E}) \leq \sum_{i \in V_c} \Pr(i \text{ is incorrectly de-anonymized}) \tag{116}$$

$$\leq \sum_{i \in V_c} 2\exp(-\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps}\Lambda) \tag{117}$$

$$= 2(n-\epsilon n-\Lambda)\exp(-\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps}\Lambda) \tag{118}$$

$$\leq 2(n-\epsilon n-\Lambda)\exp(-2\ln n - 2\ln(2(n-\epsilon n-\Lambda))) \tag{119}$$

$$= \frac{1}{n^2}. \tag{120}$$

Hence, it a.a.s. that $\Pr(\mathbf{E}) \sim 0$ as $n \to \infty$, i.e., it a.a.s. that G^a is $(1-\epsilon)$-de-anonymizable if $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{2\ln n + \ln(2(n-\epsilon n-\Lambda))}{\Lambda}$.

Second, we prove if $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \frac{(k+2)\ln n + \ln(2(n-\epsilon n-\Lambda))}{k(n-k/2-1)}$, G^a is also $(1-\epsilon)$-de-anonymizable. The proof is similar to the part 2 of Theorem 12. The difference is that we do not have to distinguish σ_0 and σ_k when $k \leq \epsilon n$. Let $\mathbf{E}$ now be the event that *there is some DA scheme* σ_k *such that* $\sigma_k \neq \sigma_0$, $\Delta_{\sigma_k} \leq \Delta_{\sigma_0}$, *and* $k > \epsilon n$. Then, we have

$$\Pr(\mathbf{E}) = \bigcup_{k=\epsilon n+1}^{n-\Lambda} \Pr(\Delta_{\sigma_k} \leq \Delta_{\sigma_0}) \tag{121}$$

$$\leq \sum_{k=\epsilon n+1}^{n-\Lambda} n^k \cdot 2\exp(-\frac{1}{4}\frac{ps^3(1-p)^2}{2-s-ps}(m_k - k/2)) \tag{122}$$

$$\leq \sum_{k=\epsilon n+1}^{n-\Lambda} n^k \cdot 2\exp(-(k+2)\ln n - \ln(2(n-\epsilon n-\Lambda))) \tag{123}$$

$$= \frac{1}{n^2}. \tag{124}$$

Consequently, it a.a.s. that $\Pr(\mathbf{E}) \sim 0$ as $n \to 0$, i.e., G^a is $(1-\epsilon)$-de-anonymizable.

In summary, G^a is $(1-\epsilon)$-de-anonymizable if $\frac{1}{4} \cdot \frac{ps^3(1-p)^2}{2-s-ps} \geq \min\{\frac{2\ln n + \ln(2(n-\epsilon n-\Lambda))}{\Lambda}, \frac{(k+2)\ln n + \ln(2(n-\epsilon n-\Lambda))}{k(n-k/2-1)}\}$. □

4.4 Quantification in General Scenarios

Although the ER model is suitable to enable elegant theoretical analysis on the de-anonymizability of graph data, the fact is that it is extremely rare, if not impossible, to see real world graph data actually follow the ER model [107]. Nevertheless, the analysis under the ER model can shed light on the theoretical quantification of the de-anonymizability of graph data in general scenarios.

In this section, we quantify the de-anonymizability of G^a in general scenarios, i.e., unlike in Section 4.3, we assume $G(V, E)$ now could be some graph following an arbitrary network model. To accelerate the quantification, we make some definitions as follows. Given a graph $G(V, E)$ with $|V| = n$ and $|E| = m$, its *graph density* is defined as $\rho = \frac{2m}{n(n-1)}$. Let $U \subseteq V$. The *subgraph* of G on U is defined as $G[U] = G(U, E_U = \{e_{i,j} \in E | i, j \in U\})$. Furthermore, let $n_U = |U|$ and $m_U = |E_U|$. Then, the *subgraph density* of G on U is $\rho_U = \frac{2m_U}{n_U(n_U-1)}$. Let U and W be two disjoint subsets of V ($U \cap W = \emptyset$), $E_{U,W} = \{e_{i,j} \in E | i \in U, j \in W\}$ be the set of edges connecting U and W, and $m_{U,W} = |E_{U,W}|$. Then, the *connectivity* between U and W is defined as $\gamma_{U,W} = \frac{m_{U,W}}{n_U \cdot n_W}$. Finally, for the seed mapping set $\mathcal{S}$, we assume it is randomly identified, which implies each user in V is selected with a probability of $q = \frac{\Lambda}{n}$. For the seed users in V, we denote them as a set S for convenience, i.e., $S = \mathcal{S}^a = \mathcal{S}^u$. For the other users, we denote them by set $A = V \setminus S$.

4.4.1 $\mathcal{S}$ based Quantification

In this subsection, we quantify the de-anonymizability of a graph given a seed mapping set $\mathcal{S}$. First, we show the condition for perfectly de-anonymizing an anonymized user in Theorem 15.

Theorem 15. *If* $\frac{1}{4} \cdot \frac{qs^3(1-\gamma_{S,A})^2}{2-s-s\gamma_{S,A}} \geq \frac{2\ln n+1}{d_i}$, *where* $q = \Lambda/n$ *and* $\gamma_{S,A} = \frac{m_{S,A}}{\Lambda(n-\Lambda)}$, *it is a.a.s. that* $\forall i \in A$, *i is perfectly de-anonymizable.*

Proof: To prove this theorem, it is sufficient to prove that $\forall i \in A$, $\Delta_{\sigma:(i,i)}(\mathcal{S}) < \Delta_{\sigma:(i,j\neq i)}(\mathcal{S})$ under any given σ (it follows that i is perfectly de-anonymizable in terms of $\Delta_{\sigma:(i,\sigma(i))}(\mathcal{S})$). Let $X = \Delta_{\sigma:(i,j\neq i)}(\mathcal{S})$ and $Y = \Delta_{\sigma:(i,i)}(\mathcal{S})$ be two random variables. Similar as in Theorem 10, we have

$$X \underset{\text{stochastically}}{\sim} \mathbf{B}(d_i q, 2s(1 - s\gamma_{S,A})) \tag{125}$$

$$Y \sim \mathbf{B}(d_i q, 2s(1 - s)). \tag{126}$$

Applying Lemma 1, we have

$$\Pr(X \leq Y) \leq 2\exp(-\frac{(\lambda_X - \lambda_Y)^2}{8(\lambda_X + \lambda_Y)}) \tag{127}$$

$$= 2\exp(-\frac{1}{4}\frac{qs^3(1 - \gamma_{S,A})^2}{2 - s - s\gamma_{S,A}} d_i) \tag{128}$$

$$\leq \frac{1}{n^2}. \tag{129}$$

Since $\sum\limits_{n>0} \frac{1}{n^2} = \frac{\pi^2}{6} < \infty$, it is a.a.s. that $\Pr(X \leq Y) \sim 0$ as $n \to \infty$, i.e., i is perfectly de-anonymizable. □

In Theorem 15, the condition where a user is perfectly de-anonymized is quantified. We further quantify the condition to perfectly de-anonymize all the users in A in Theorem 16.

Theorem 16. *If* $\frac{1}{4} \cdot \frac{qs^3(1-\gamma_{S,A})^2}{2-s-s\gamma_{S,A}} \geq \frac{2\ln n + \ln(2(n-\Lambda))}{d_i}$, *where* $q = \Lambda/n$ *and* $\gamma_{S,A} = \frac{m_{S,A}}{\Lambda(n-\Lambda)}$, *it is a.a.s. that* G^a *is perfectly de-anonymizable.*

Proof: This theorem can be proven by using similar techniques as in Theorems 11 and 15. □

4.4.2 Sophisticated Quantification: Considering more Structure Information

In the previous subsection, the perfect de-anonymizability of graph data is quantified in general scenarios based on $\mathcal{S}$. As we discussed in Section 4.3, for $i \in A$, besides the structural connection to the users in S, the structural information between i and other

users in A is also helpful to improve the DA performance (as shown in Theorem 12). Similar to the quantification under the ER model, we quantify the de-anonymizability of graph data by considering the overall structure information in Theorem 17.

Theorem 17. *(i) If* $\frac{1}{4}\cdot\frac{s^3(1-\max\{\gamma_{V_0,V_k},\rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k},\rho_{V_k}\}}\geq\frac{2\ln n+1}{m_{V_0,V_k}+m_{V_k}-k/2}$*, it is a.a.s. that* $\Delta_{\sigma_0}<\Delta_{\sigma_k}$ $(k\in[2,n])$*, i.e., it is a.a.s. that the perfect DA scheme* σ_0 *induces less edge difference than any given DA scheme* $\sigma_k\neq\sigma_0$*; (ii) If* $\frac{1}{4}\cdot\frac{s^3(1-\max\{\gamma_{V_0,V_k},\rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k},\rho_{V_k}\}}\geq\frac{(k+2)\ln n+\ln(2(n-\Lambda-1))}{m_{V_0,V_k}+m_{V_k}-k/2}$*, it is a.a.s. that the perfect DA scheme* σ_0 *is the only scheme inducing the least edge difference, i.e.,* G^a *is perfectly de-anonymizable.*

Proof: (i) Let $V_k\subseteq V\setminus S$ be the set of incorrectly de-anonymized users under $\sigma_k\neq\sigma_0$ and $V_0=V\setminus V_k$. Furthermore, let $X=\Delta_{\sigma_k}$ and $Y=\Delta_{\sigma_0}$ be two random variables. Then, similar as the derivation in Theorem 12, we have

$$Y\sim\mathbf{B}(m,2s(1-s)). \tag{130}$$

Furthermore, we can consider four cases to quantify X. First, the edge difference caused by the edges in E_{V_0} follows $\mathbf{B}(m_{V_0},2s(1-s))$; second, the edge difference caused by the edges in E_{V_0,V_k} stochastically follows $\mathbf{B}(m_{V_0,V_k},2s(1-s\gamma_{V_0,V_k}))$; third, the edge difference caused by the non-transposition edges in E_k stochastically follows $\mathbf{B}(m_{V_k}-x,2s(1-s\rho_{V_k}))$, where x here is the number of transposition edges under σ_k; and finally, the edge difference caused by the transposition edges in E_k follows $\mathbf{B}(x,2s(1-s))$. Since $x\leq k/2$, we have

$$\begin{aligned}X &\underset{\text{stochastically}}{\geq} \mathbf{B}(m_{V_0},2s(1-s))+\mathbf{B}(m_{V_0,V_k},2s(1-s\gamma_{V_0,V_k})) &(131)\\ &\quad+\mathbf{B}(m_{V_k}-k/2,2s(1-s\rho_{V_k}))+\mathbf{B}(k/2,2s(1-s))\\ &\geq \mathbf{B}(m_{V_0},2s(1-s))+\mathbf{B}(m_{V_0,V_k}+m_{V_k}-k/2, &(132)\\ &\quad 2s(1-s\cdot\max\{\gamma_{V_0,V_k},\rho_{V_k}\}))\\ &\quad+\mathbf{B}(k/2,2s(1-s)).\end{aligned}$$

Define $\widetilde{X} \sim \mathbf{B}(m_{V_0,V_k} + m_{V_k} - k/2, 2s(1 - s \cdot \max\{\gamma_{V_0,V_k}, \rho_{V_k}\}))$ and $\widetilde{Y} \sim \mathbf{B}(m_{V_0,V_k} + m_{V_k} - k/2, 2s(1-s))$. Then, we have

$$\begin{aligned}
\Pr(X \leq Y) &\underset{\text{stochastically}}{=} \Pr(\widetilde{X} \leq \widetilde{Y}) && (133)\\
&\leq 2\exp(-\frac{1}{4}\frac{s^3(1-\max\{\gamma_{V_0,V_k}, \rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k}, \rho_{V_k}\}} \\
&\quad (m_{V_0,V_k} + m_{V_k} - k/2)) && (134)\\
&\leq 2\exp(-2\ln n - 1) && (135)\\
&\leq \frac{1}{n^2}. && (136)
\end{aligned}$$

Since $\sum\limits_{n>0} \frac{1}{n^2} = \frac{\pi^2}{6} < \infty$, it is a.a.s. that $\Pr(X \leq Y) \sim 0$ as $n \to \infty$, i.e., it is a.a.s. that $\Delta_{\sigma_0} < \Delta_{\sigma_k}$ given $\sigma_k \neq \sigma_0$.

(ii) Similar as in the proof Theorem 12, let $\mathbf{E}$ be the event that *there exists some* $\sigma_k \neq \sigma_0$ *and* $\Delta_{\sigma_k} \leq \Delta_{\sigma_0}$. Then, based on the union bound, we have

$$\begin{aligned}
\Pr(\mathbf{E}) &= \bigcup_{k=2}^{n-\Lambda} \Pr(\Delta_{\sigma_k} \leq \Delta_{\sigma_0}) && (137)\\
&\leq \sum_{k=2}^{n-\Lambda} n^k \cdot \Pr(\Delta_{\sigma_k} \leq \Delta_{\sigma_0}) && (138)\\
&\leq \sum_{k=2}^{n-\Lambda} n^k \cdot 2\exp(-\frac{1}{4}\frac{s^3(1-\max\{\gamma_{V_0,V_k}, \rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k}, \rho_{V_k}\}} \\
&\quad (m_{V_0,V_k} + m_{V_k} - k/2)) && (139)\\
&\leq \sum_{k=2}^{n-\Lambda} n^k \cdot 2\exp(-(k+2)\ln n - \ln(2(n-\Lambda-1))) && (140)\\
&= \frac{1}{n^2}. && (141)
\end{aligned}$$

Consequently, it is a.a.s. that σ_0 is the only scheme inducing the least edge difference, i.e., it is a.a.s. that G^a is perfectly de-anonymizable. □

Similar as Theorem 12, Theorem 17 also implies a large scale graph is perfectly de-anonymizable without seed information in general scenarios. We summarize the condition in Corollary 2.

Corollary 2. *If* $\frac{1}{4} \cdot \frac{s^3(1-\max\{\gamma_{V_0,V_k},\rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k},\rho_{V_k}\}} \geq \frac{(k+2)\ln n+\ln(2(n-1))}{m_{V_0,V_k}+m_{V_k}-k/2}$*, it is a.a.s. that the perfect DA scheme* σ_0 *is the only scheme inducing the least edge difference, i.e.,* G^a *is perfectly de-anonymizable.*

Based on Theorems 16, 17 and Corollary 2, it is straightforward to have the following conclusion.

Theorem 18. *If* $\frac{1}{4} \cdot \frac{qs^3(1-\gamma_{S,A})^2}{2-s-s\gamma_{S,A}} \geq \frac{2\ln n+\ln(2(n-\Lambda))}{d_i}$ *or* $\frac{1}{4} \cdot \frac{s^3(1-\max\{\gamma_{V_0,V_k},\rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k},\rho_{V_k}\}} \geq \frac{(k+2)\ln n+\ln(2(n-\Lambda-1))}{m_{V_0,V_k}+m_{V_k}-k/2}$*, where* $\Lambda \in [0,n]$*, it is a.a.s. that* G^a *is perfectly de-anonymizable.*

4.4.3 Quantification with Error Toleration

Now, we quantify the $(1-\epsilon)$-*de-anonymizability* of graph data in general scenarios, where now ϵn ($\epsilon \in [0, 1-\frac{\Lambda}{n}]$) users are allowed to be incorrectly de-anonymized. We demonstrate the quantification in Theorem 19.

Theorem 19. *If (i)* $\frac{1}{4}\cdot\frac{qs^3(1-\gamma_{S,A})^2}{2-s-s\gamma_{S,A}} \geq \frac{2\ln n+\ln(2(n-\epsilon n-\Lambda))}{d_i}$ *or (ii)* $\frac{1}{4}\cdot\frac{s^3(1-\max\{\gamma_{V_0,V_k},\rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k},\rho_{V_k}\}} \geq \frac{(k+2)\ln n+\ln(2(n-\epsilon n-\Lambda))}{m_{V_0,V_k}+m_{V_k}-k/2}$*, where* $\Lambda \in [0,n]$*,* G^a *is* $(1-\epsilon)$*-de-anonymizable.*

Proof: (i) As in Theorem 14, let V_c be the set of users that are perfectly de-anonymizable and $|V_c| = n - \epsilon n - \Lambda$. Furthermore, let $\mathbf{E}$ be the event that *there exists at least one incorrectly de-anonymizable user in* V_c. Then, we have

$$\Pr(\mathbf{E}) \leq \sum_{i\in V_c} Pr(i \text{ is incorrectly de-anonymized}) \tag{142}$$

$$\leq \sum_{i\in V_c} 2\exp(-\frac{1}{4}\frac{qs^3(1-\gamma_{S,A})^2}{2-s-s\gamma_{S,A}}d_i) \tag{143}$$

$$\leq \sum_{i\in V_c} 2\exp(-2\ln n - \ln(2(n-\epsilon n-\Lambda))) \tag{144}$$

$$= \frac{1}{n^2}. \tag{145}$$

Consequently, it is a.a.s. that $\Pr(\mathbf{E}) \sim 0$ as $n \to \infty$, i.e. it is a.a.s. that G^a is $(1-\epsilon)$-de-anonymizable.

(*ii*) As in Theorem 14, we do not have to distinguish σ_k with σ_0 when $k \leq \epsilon n$. Let $\mathbf{E}$ be the event that *there exists some* σ_k *such that* $\Delta_{\sigma_k} \leq \Delta_{\sigma_0}$ *and* $k > \epsilon n$. Then, we have

$$\Pr(\mathbf{E}) \leq \sum_{k=\epsilon n+1}^{n-\Lambda} n^k \cdot \Pr(\Delta_{\sigma_k} \leq \Delta_{\sigma_0}) \tag{146}$$

$$\leq \sum_{k=\epsilon n+1}^{n-\Lambda} n^k \cdot 2\exp(-\frac{1}{4}\frac{s^3(1-\max\{\gamma_{V_0,V_k},\rho_{V_k}\})^2}{2-s-s\cdot\max\{\gamma_{V_0,V_k},\rho_{V_k}\}}$$

$$(m_{V_0,V_k}+m_{V_k}-k/2)) \tag{147}$$

$$\leq \sum_{k=\epsilon n+1}^{n-\Lambda} n^k \cdot 2\exp(-(k+2)\ln n - \ln(2(n-\epsilon n-\Lambda))) \tag{148}$$

$$= \frac{1}{n^2}. \tag{149}$$

Again, it is a.a.s. that $\Pr(\mathbf{E}) \sim 0$ as $n \to \infty$, i.e. it is a.a.s. that G^a is $(1-\epsilon)$-de-anonymizable. □

4.5 Large Scale Evaluation

4.5.1 Datasets and Setup

In the evaluation, we employ 24 various real world social datasets that mainly come from the *Stanford Large Network Dataset Collection* [16], *ASU Social Computing Data Repository* [1], and other sources [15, 135]. The employed datasets are shown in Table 9 with preliminary statistics, where n is the number of users (nodes), m is the number of edges among users, ρ is the graph density, $\overline{d}$ is the average degree of the users, and $p(k)$ $(k = 1, 5, 10)$ is the percentage of users with degree less than or equal to k. We further briefly introduce the datasets below. The detailed descriptions can be found in the corresponding references.

- **Hyves** [1]. Hyves is the most popular social network in the Netherlands and competes in that country with other well known international social networks

Table 9: Dataset statistics.

Name	n	m	ρ	$\overline{d}$	$p(1)$	$p(5)$	$p(10)$
Hyves	1,402,673	2,777,419	2.82E-06	3.96	56.76%	88.74%	91.80%
Douban	154,908	327,162	2.73E-05	4.22	66.57%	90.81%	93.86%
Friendster	5,689,498	14,067,887	8.69E-07	4.95	60.19%	91.27%	95.86%
YouTube	1,138,499	2,990,443	4.61E-06	5.25	53.16%	85.53%	92.78%
Flixster	2,523,386	7,918,801	2.49E-06	6.28	59.49%	87.26%	92.86%
Last.fm	1,191,812	4,519,340	6.36E-06	7.58	47.27%	81.62%	89.54%
FB-NO-wall	45,813	183,412	1.75E-04	8.01	24.18%	60.91%	77.42%
Gowalla	196,591	950,327	4.92E-05	9.70	25.20%	64.50%	79.90%
Foursquare	639,014	3,214,986	1.57E-05	10.06	51.10%	79.11%	83.21%
Enron	33,696	180,811	3.19E-04	10.73	28.09%	67.86%	82.88%
Skitter	1,694,616	11,094,209	7.73E-06	13.09	12.80%	55.41%	76.21%
Slashdot	82,168	582,533	1.73E-04	14.18	2.19%	64.78%	78.30%
Digg	771,229	5,907,413	1.99E-05	15.32	45.64%	77.31%	85.97%
LiveJournal	4,843,953	43,362,750	3.70E-06	17.90	20.99%	50.53%	64.88%
HepPh	11,204	117,649	1.87E-03	21.00	9.95%	49.99%	66.45%
AstroPh	17,903	197,031	1.23E-03	22.01	5.34%	33.69%	50.66%
FB-NO-links	63,731	817,090	4.02E-04	25.64	12.71%	36.11%	50.02%
Pokec	1,632,803	22,301,964	1.67E-05	27.32	10.04%	30.66%	44.48%
BlogCatalog	97,884	1,668,647	3.48E-04	34.10	28.24%	59.59%	71.45%
Google+	4,692,671	90,751,480	8.24E-06	38.68	5.44%	27.33%	46.37%
Livemocha	104,103	2,193,083	4.05E-04	42.13	6.56%	27.56%	44.02%
Twitter	456,293	12,508,272	1.20E-04	54.83	5.30%	19.76%	34.50%
Orkut	3,072,441	117,185,083	2.48E-05	76.28	2.21%	7.28%	13.35%
Flickr	80,513	5,899,882	1.82E-03	146.56	0.00%	11.63%	20.58%

(e.g., Facebook, MySpace) [1]. The employed Hyves dataset is a friendship network.

- **Douban** [1]. Douban is a Chinese Web 2.0 site that provides user review and recommendation services for movies, books, and music. It is also the largest online Chinese language book, movie, and music database and one of the largest online communities in China. The employed Douban dataset is a friendship network of the users of Douban.

- **Friendster** [1]. Friendster is a social gaming site and was a social networking service website before being redesigned. The service allows users to contact other members, maintain those contacts, and share online content and media with those contacts. The employed Friendster dataset is a friendship network.

- **YouTube** [1]. YouTube is a well known video sharing website on which users can upload, share, and view videos. The employed YouTube dataset is a user contact network.

- **Flixster** [1]. Flixster is a social movie site allowing users to share movie ratings, discover new movies and meet others with similar tastes in movies. The employed Flixster is a friendship network.

- **Last.fm** [1]. Last.fm is a music discovery service that gives you personalized recommendations based on the music you listen to. The employed Last.fm dataset is a friendship network.

- **Facebook-New Orleans-links (FB-NO-links) and Facebook-New Orleans-wall (FB-NO-wall)** [135]. Facebook is one of the most popular social networks, which connects people with friends and others who work, study, and live around them. The employed FB-NO-links dataset is a Facebook friendship

network at the New Orleans area and the FB-NO-wall is a Facebook interaction (wall posts) network at the New Orleans area.

- **Gowalla** [16]. Gowalla is a location-based social networking website where users share their locations by checking-in. The employed Gowalla dataset is a friendship network.

- **Foursquare** [1]. Foursquare helps people to find the places to go with friends and discover food, nightlife, and entertainment for users. The employed Foursquare dataset is friendship network.

- **Enron** [16]. Enron is an email communication dataset released by Federal Energy Regulatory Commission during its investigation.

- **Skitter** [16]. The Skitter dataset is an Internet topology graph of *Autonomous Systems*.

- **Slashdot** [16]. Slashdot is a technology-related news website known for its specific user community. The website features user-submitted and editor-evaluated current primarily technology oriented news. The employed Slashdot dataset is a friendship network.

- **Digg** [1]. Digg is a news aggregator with an editorially driven front page, aiming to select stories specifically for the Internet audience such as science, trending political issues, and viral Internet issues. The employed Digg dataset is a friendship network.

- **LiveJournal** [16]. LiveJournal is social network for journals and blogs. It also offers privacy controls, photo storage, publishing tools, style templates, and online communities for many interests. The employed LiveJournal dataset is a friendship network.

- **HepPh** [16]. HepPh is a citation graph of the papers posted on arXiv in the high-energy physics area.

- **AstroPh** [16]. AstroPh is a collaboration network of the authors of papers posted on arXiv in the astro physics area.

- **Pokec** [16]. Pokec is the most popular on-line social network in Slovakia. The employed Pokec dataset is a friendship network.

- **BlogCatalog** [1]. BlogCatalog is a social blog directory which manages the bloggers and their blogs. The employed BlogCatalog dataset is a friendship network.

- **Google+** [15]. Google+ is one of the most popular social networking and identity services. The employed Google+ dataset is a friendship network.

- **Livemocha** [1]. Livemocha is the world's largest online language learning community, offering free and paid online language courses in 35 languages. The employed Livemocha dataset is a friendship network.

- **Twitter** [1]. Twitter is an online social networking and microblogging service that enables users to send and read short 140-character text messages, called "tweets". The employed Twitter dataset is a friendship network.

- **Orkut** [1]. Orkut is an on-line social network where users form friendship with each other. It also allows users to form a group which other members can then join. The employed Orkut dataset is a friendship network.

- **Flickr** [1]. Flickr is an image hosting and video hosting website. The employed dataset is a friendship network of Flickr users.

For each employed dataset, we use the raw data except for removing isolated users (most datasets do not contain any isolated users). Note that, our quantification is

not limited to connected graphs. It is also applicable to disconnected social networks. Furthermore, we do not consider the direction information even if a dataset is a directed network. Again, this assumption does not limit the evaluation or quantification. Since the *direction information* can be used to improve the effectiveness of DA attacks [104], it is possible that our quantification and evaluation can be improved if we have more knowledge, e.g., the direction information. One of the future works is to quantify the de-anonymizability of directed social networks.

To generate the anonymized and auxiliary graphs, we follow the data sampling approach discussed in Section 4.2, i.e., we construct G^a and G^u from the raw data using the sampling probabilities s_a and s_u, respectively. Here, for simplicity, we set $s_a = s_u = s$. After constructing G^a and G^u, the seed mappings are chosen randomly from them (note that, seed mappings are some pre-known user mappings between G^a and G^u), which implies that the high-degree users are not given preference as in [69, 151] although they may be more helpful as seed mappings. Consequently, our evaluation results represent the general results of our quantification. Each group of evaluations is repeated for 50 times and the results are the average of these 50 runs.

We quantify the de-anonymizability of a graph using seed information and using the overall structural information, respectively. Therefore, we use suffixes "-S" and "-A" to distinguish these two scenarios (e.g., Twiiter-A and Twitter-S), where "-S" and "-A" imply using seed information and overall structural information, respectively. *If we do not specify the suffix or the particular context, it implies using the overall structural information by default.*

4.5.2 Evaluation of Perfect De-anonymizability

In this subsection, we evaluate the condition on perfect de-anonymizability of the datasets in Table 9.

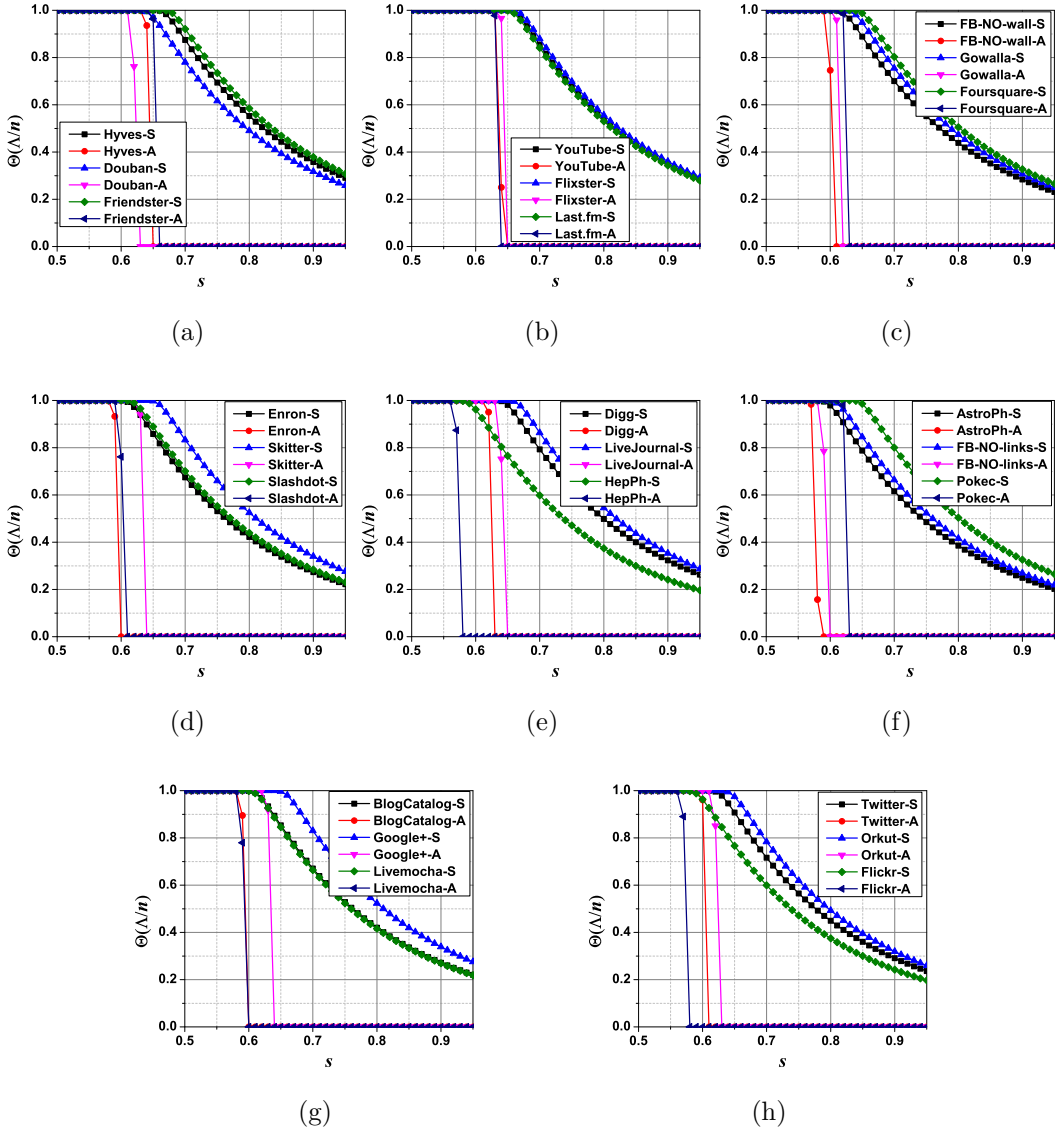


Figure 8: Perfect DA: $\Theta(\Lambda/n)$ vs. s. Since the quantification (Theorem 18) for perfect DA is meaningful for large n, we set $n = 1000/\rho$ for each social network in this group of evaluations. All the other network properties, e.g., ρ, $\overline{d}$, degree distribution, etc., remain the same as in the original dataset.

4.5.2.1 Evaluation on Λ

Based on our quantification, we evaluate the requirements on the size of seed mappings Λ and the sampling rate s for the perfect de-anonymizability of each dataset in Fig.8. Since all the datasets have different sizes, for convenience, we show $\Theta(\Lambda/n)$ instead of Λ directly.

From Fig.8, we have the following observations.

• If the overall structural information is considered, each dataset is *asymptotically perfectly de-anonymizable*[6] even without any seed information when s is above some threshold value, which is consistent with our theoretical quantification[7]. For instance, the Twitter dataset is asymptotically perfectly de-anonymizable when $s \geq 0.61$ without seed information if the overall structural information is considered. This implies that the structure itself is sufficient to break the privacy. The reason for this result is that the perfect DA scheme induces the least edge difference as shown in our quantification.

• If s is below some threshold value, it is necessary to have $\Theta(n)$ seed mappings to perfectly de-anonymize each social network, i.e., each social network is not perfectly de-anonymizable unless $\Theta(n)$ users are identified as seeds. For instance, Google+ is not perfectly de-anonymizable when $s < 0.61$. The reason is that a small s implies less edges are sampled into G^a and G^u. It follows that most of the users are low degree users and thus the structural information is not sufficient to achieve perfect DA.

• For the DA only based on seed information ("*-S"), to achieve perfect DA, the required number of seed mappings decreases when s increases as expected. For instance, to perfectly de-anonymize Google+, 49.27% seed users are needed when $s = 0.8$ while 31.89% seed users are needed when $s = 0.9$. This is because a large s implies more structural similarity between G^a and G^u. Thus, less seed mappings are needed to distinguish all the users.

• Given some s, a social network with higher graph density requires fewer seed mappings. For example, to be perfectly de-anonymizable, 49.27% ($\rho = 2.48\text{E-}5$) seed

[6]To be accurately, *asymptotically perfectly de-anonymizable* here implies $\Theta(n)$ users of each dataset can be successfully de-anonymized.

[7]Actually, the quantification does not implies a computationally efficient algorithm. It is still an open problem to find an efficient (polynomial-time) algorithm with provable performance guarantee.

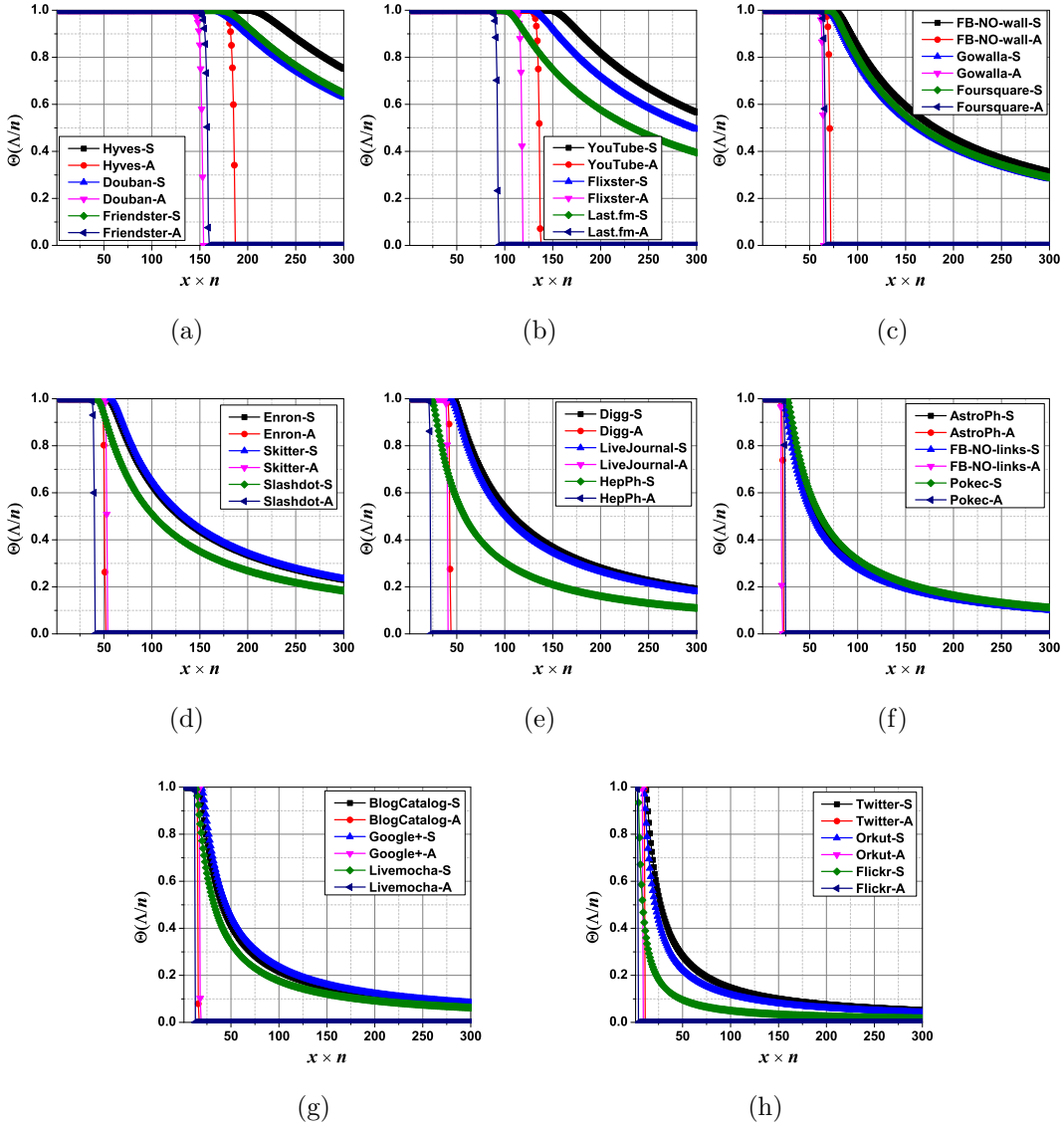


Figure 9: Perfect DA: $\Theta(\Lambda/n)$ vs. n. Default setting: $s = 0.7$.

users are required for Orkut-S while 37.47% seed users are required for Flickr-S ($\rho =$ 1.82E-3). This is also true for the overall structural information based DA. This is because a higher graph density implies that more structural information is carried by the data, followed by more structural information can be used to distinguish users.

Now, we examine the behavior of Λ when we fix the graph density of each social network while varying n. The results are shown in Fig.9, where $x \times n$ (the x-axis) denotes the number of users is x times of the original size n.

From Fig.9, we have the following observations.

• When n is above some threshold value, each social network is perfectly de-anonymizable based on the overall structural information, which confirms the conclusion of Theorem 17. The reason is straightforward. More structural information will be available when n increases and ρ is fixed. Consequently, more users can be de-anonymized based on the structural information. Because of the same reason, for the seed-based DA, the required number of seed mappings decreases when n increases.

• Given $\Theta(\Lambda/n)$, the required threshold value on n is smaller for social networks with high graph densities and vice versa. This is because a high ρ implies more structural information is available followed by more similarity between G^a and G^u when s is fixed.

• Similar to the scenario of changing s, when n is below some threshold value, it is necessary to have $\Theta(n)$ seed user mappings to perfectly de-anonymize a social network. This can be seen from our quantification: it is a.a.s. that σ_0 induces the least edge difference when the required condition holds and $n \to \infty$, i.e., n should be large enough.

4.5.2.2 Evaluation on n

In this subsection, we study the condition on n to perfectly de-anonymize a social network given different s or Λ. The objective of this group of evaluation is to study the asymptotic behavior of n in different scenarios, since our quantification is mathematically meaningful when n is a large number. Furthermore, based on our quantification, when $n \to \infty$, the overall structure based quantification will dominate the perfect de-anonymizability of a social network (this claim can be confirmed by the evaluation results in Fig.11). Consequently, we consider the overall structural information (including seed mappings) in the evaluation of n. When s is changed from 0.5 to 0.95, the requirement on the lower bound of n, i.e., $\Omega(n)$, is shown in

Fig.10. From Fig.10, we have two observations as follows.

• When s increases, $\Omega(n)$ decreases, e.g., to perfectly de-anonymize Google+, $\Omega(n)$ decreases from 2.85E8 to 3.19E7 when s increases from 0.5 to 0.9. This is because a large s implies more similarity between G^a and G^u since they share more common edges. Consequently, the condition on $\Omega(n)$ to perfectly de-anonymize a social network becomes loose. This observation can also be explained by our quantification. From Theorem 18, a large s implies a large value on the left hand of each condition, followed by smaller n requirement on the right hand.

• The graph density has positive influence on $\Omega(n)$, i.e., a social network with high graph density requires loose condition on $\Omega(n)$ for perfect DA. For instance, Orkut ($\rho = 2.48\text{E-}5$) requires a smaller $\Omega(n)$ than Google+ ($\rho = 8.24\text{E-}6$). The reason is that a large ρ implies more structural information is carried by the dataset. Therefore, it is much easier to perfectly de-anonymize this dataset.

Now, we want to study the impact of Λ on $\Omega(n)$ to perfectly de-anonymize a social network. The results are shown in Fig.11. From Fig.11, we have the following two observations.

• When Ω (i.e., $\Theta(\Omega/n)$) increases, $\Omega(n)$ only has a very slight decrease, e.g., the $\Omega(n)$ of Friendster, Skitter, LiveJournal, etc. This is because given that $\Omega(n)$, the overall structural information based DA has already been achieved. However, even the seed information based DA can de-anonymize a large portion of each social network given the same $\Omega(n)$ (as shown in Fig.8 and Fig.9), more seed mappings are necessary to perfectly de-anonymize all the users.

• Again, the graph density has positive influence on $\Omega(n)$ in different settings of Ω. The reason is the same as explained before: a large ρ implies more similarity between G^a and G^u when s is fixed.

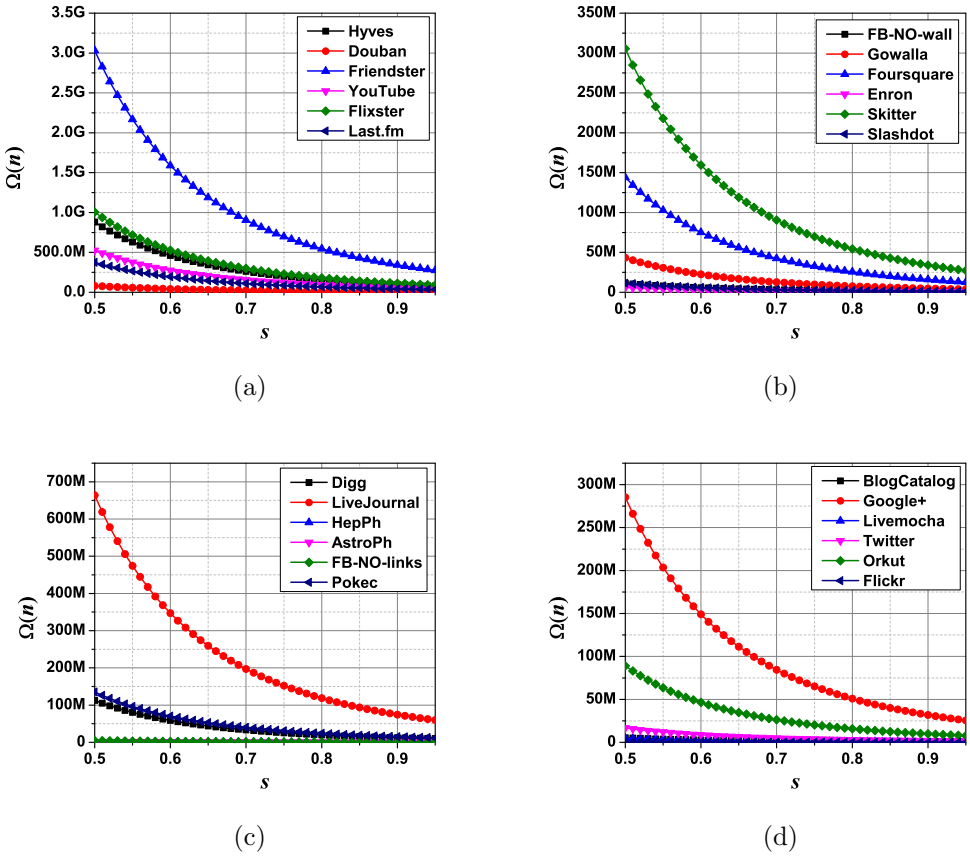


Figure 10: Perfect DA: n vs. s. Default setting: $\Lambda/n = 0.015$ (1.5% users are randomly chosen as seed mappings).

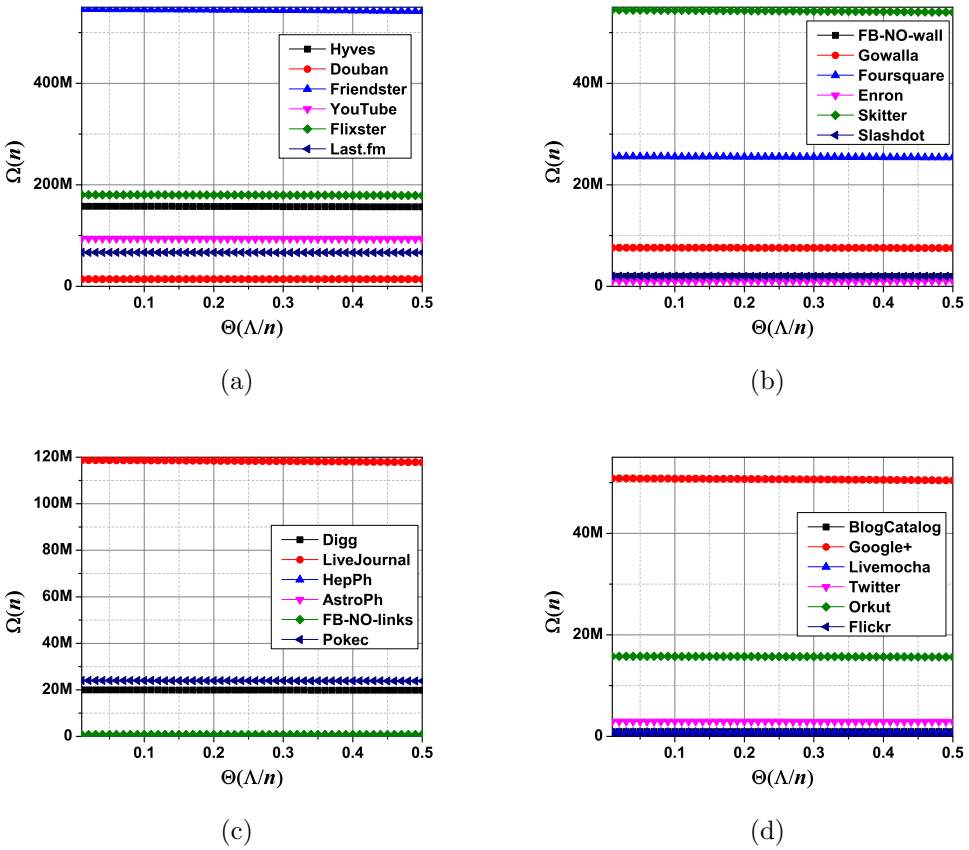


Figure 11: Perfect DA: n vs. Λ. Default setting: $s = 0.8$.

4.5.3 Evaluation of $(1-\epsilon)$-De-anonymizability

4.5.3.1 Evaluation on $(1-\epsilon)$

In this subsection, we evaluate the actual de-anonymizability of the 24 real world datasets by quantitatively demonstrating $(1-\epsilon)$ (note that, ϵn is the error tolerated during the DA process), i.e., *how many users in each social network can be successfully de-anonymized in each specific scenario.*

When all the structural information (including seed mappings) are considered, *the lower bound on the percentage of de-anonymizable users in the 24 social networks*, i.e., $\Omega(1-\epsilon)$, is shown in Fig.12 with different s. From Fig.12, we have the following observations.

• All the 24 social networks are partially de-anonymizable although they may not be perfectly de-anonymizable. For instance, when $s = 0.55$, 20.88% YouTube users, 33.62% Foursquare users, 66.69% Facebook users at New Orleans, 72.94% Google+ users, and 97.6% Twitter users are de-anonymizable based on the overall structural information. Consequently, the obtained quantitative results confirmed the success of existing heuristic algorithms [104, 127]. This is also consistent with our quantification on $(1-\epsilon)$-DA: *if the low-degree users are treated as the tolerated DA errors, the high-degree users are more likely to be successfully de-anonymized*, i.e., these social networks are partially de-anonymizable. In other words, when perfect DA is not achievable, these high-degree users are still de-anonymizable since they carry enough structural information.

• When s increases, $\Omega(1-\epsilon)$ also increases, i.e., more users can be successfully de-anonymized for each social network. For instance, when s changes from 0.5 to 0.65, the percentage of de-anonymizable users of Google+ increases from 58.76% to 99%. The reason is similar as explained in the previous subsection: a large s implies more common edges shared by G^a and G^u, i.e., more structural similarity between G^a and G^u. Consequently, it is more likely that the correct user DA induces less edge

difference (DA error).

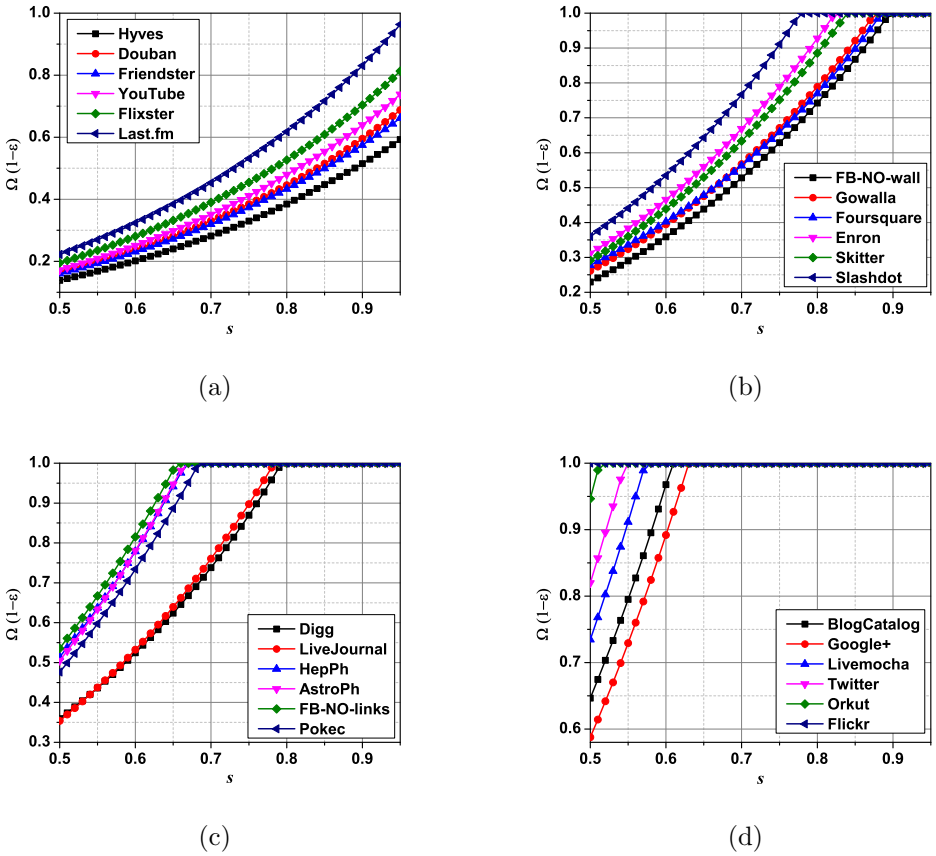
• When s is increased above some value, several social networks can be asymptotically perfectly de-anonymizable ($\Theta(n)$ users can be successfully de-anonymized). For instance, when $s \geq 0.78$, $s \geq 0.66$, and $s \geq 0.63$, over 99% users of Slashdot, FB-NO-link, and Google+ can be successfully de-anonymized, respectively. This fact comes from the same reason as the previous observation: a large s implies more structural similarity followed by more de-anonymizable a social network is.

• The social networks with higher average degree $\overline{d}$ is more de-anonymizable, e.g., when $s = 0.6$, 53.23% LiveJournal users ($\overline{d} = 17.9$) are perfectly de-anonymizable while 73.38% Pokec users ($\overline{d} = 27.32$) are perfectly de-anonymizable. The reason is evident: a higher $\overline{d}$ implies more common edges in G^a and G^u. Therefore, the correct DA is more likely inducing less edge difference.

Now, we study the $(1-\epsilon)$-de-anonymizability of the 24 social networks when we fix the network density, s, and Λ/n while change n. The results are shown in Fig.13. From Fig.13, we have the following observations.

• When n increases, the percentage of de-anonymizable users of each social network also increases for both seed-based DA and overall SDA. For instance, when the network size changes from $10n$ to $20n$, the percentage of de-anonymizable Flickr users increases from 41.65% to 59.08% in seed-based DA; similarly, when network size is $5n$, 67.81% of LiveJournal users are de-anonymizable while when the network size is above $10.5n$, LiveJournal is asymptotically perfectly de-anonymizable. This fact is consistent with our quantification. The reason is that a large n implies richer structural information when ρ is fixed. Hence, more users are de-anonymizable.

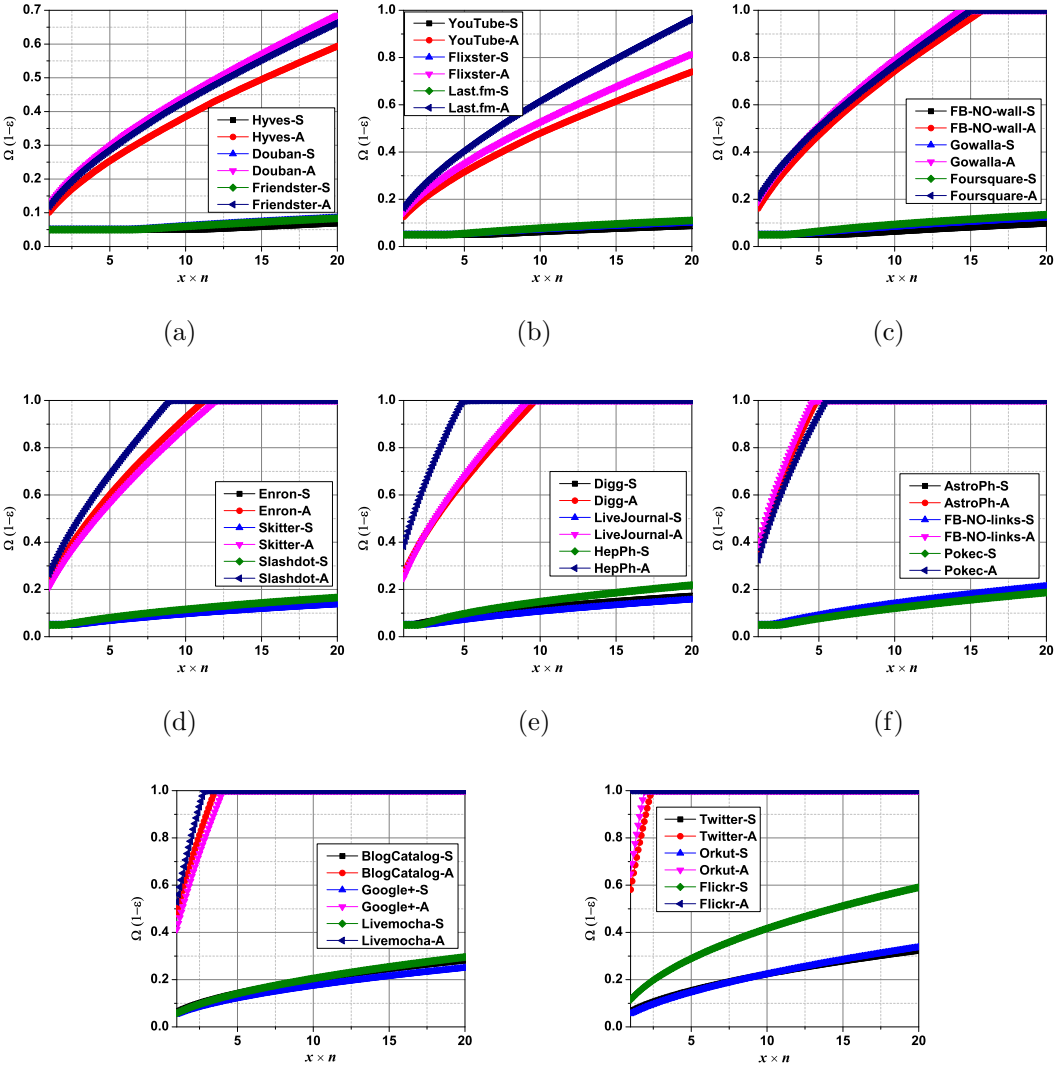
• As expected, the overall structural information is more powerful in de-anonymizing social networks. This is also consistent with our quantification. Since more structural information is considered, the probability that correct DA induces more edge differences than incorrect DA will be decreased. Consequently, "*-A" de-anonymizes



(a) (b)

(c) (d)

Figure 12: $(1-\epsilon)$-DA: $\Omega(1-\epsilon)$ vs. s. Default setting: $\Lambda = 0.05n$ (5% users are seeds).



(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 13: $(1-\epsilon)$-DA: $\Omega(1-\epsilon)$ vs. n. Default setting: $s = 0.8$ and $\Lambda/n = 0.05$.

more users than "*-S".

• As validated before, graph density also has positive impact on $\Omega(1-\epsilon)$, i.e., a social network with a high graph density is more de-anonymizable. The reason still comes from the fact that a high ρ implies more structural similarity between G^a and G^u.

Intuitively, if we have more seed mappings, more users should be de-anonymizable

even we do not consider the overall structural information. Theoretically, this intuition is quantified in Theorem 16. We evaluate this quantification by studying the impacts of the number of seed mappings on the percentage of de-anonymizable users. The results are shown in Fig.14. From Fig.14, we have the following observations.

• When more seed mappings are available, more users are de-anonymizable, e.g., when $\Omega(\Lambda/n)$ changes from 0.05 to 0.15, the percentage of de-anonymizable Google+ users increases from 40.07% to 72.28%. The reason is evident since more seed mappings implies more knowledge is available to improve the DA accuracy, which can also be seen from our quantification.

• Although ρ and $\overline{d}$ have a positive influence to $\Omega(1-\epsilon)$, it is still possible that a social network with smaller ρ or $\overline{d}$ may be more de-anonymizable than a social network with higher ρ or $\overline{d}$ in some cases, e.g., BlogCatalog has a smaller $\overline{d}$ while larger ρ than Google+, and Orkut has a smaller ρ while larger $\overline{d}$ than BlogCatalog. This is because the seed mappings in seed-based DA are randomly identified and the DA process is also affected by the degree distribution of the social network. Consequently, both ρ and $\overline{d}$ have impacts on the de-anonymizability of a social network. However, it is difficult to determine which one will dominate the de-anonymizability. Generally speaking, the richer the structural information, i.e., the higher ρ and $\overline{d}$, the more de-anonymizable the social network is.

4.5.3.2 Evaluation on Λ

In this subsection, we evaluate the condition on Λ in $(1-\epsilon)$-DA. When $\epsilon = 0.4$, i.e. up to 40% user DA error is tolerable, the condition on Λ to perfectly de-anonymize at least $1-\epsilon = 60\%$ users of each social network under different settings of s is shown in Fig.15. From Fig.15, we can observe that:

• When s is below some threshold value, $\Theta((1-\epsilon)n)$ seed mappings are necessary to perfectly de-anonymize $(1-\epsilon)n$ anonymized users. For instance, when $s < 0.72$,

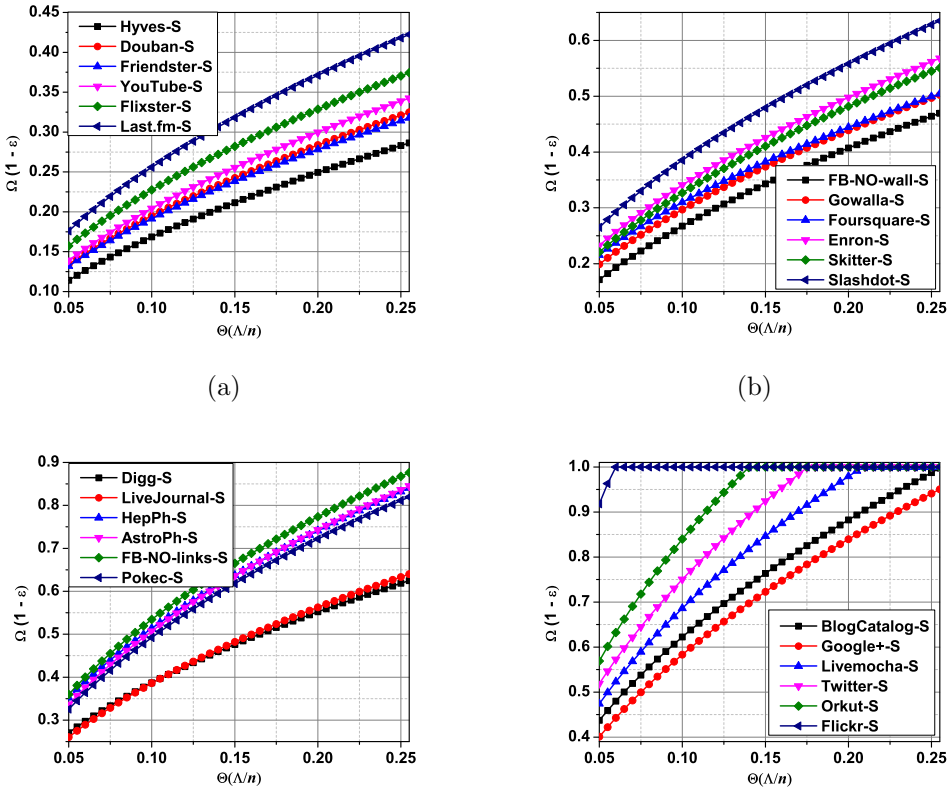


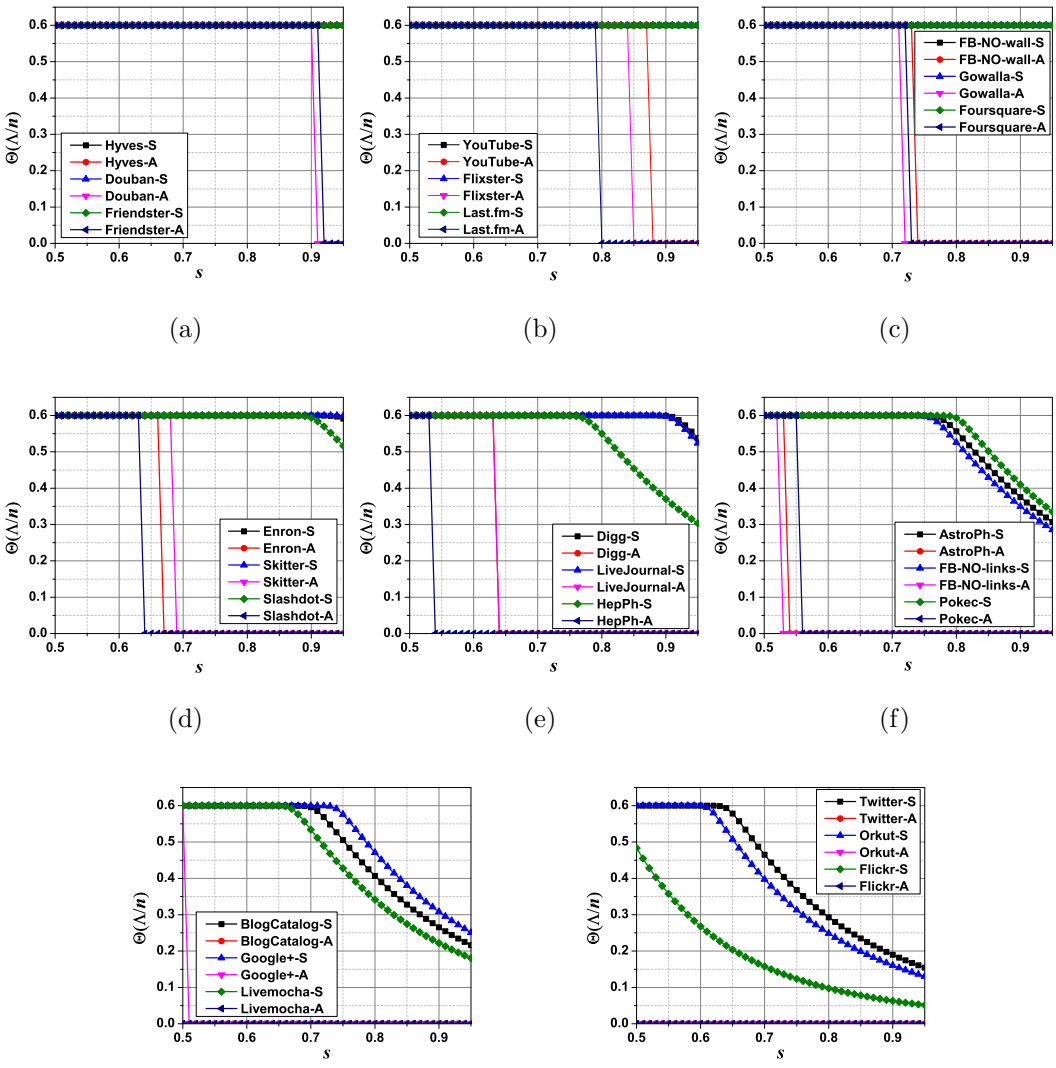
Figure 14: $(1-\epsilon)$-DA: $\Omega(1-\epsilon)$ vs. Λ. Default setting: $s = 0.8$

$\Theta(\Lambda/n) \sim 0.6$ for Google+ in seed-based DA, i.e., almost 60% Google+ users have to be identified as seeds; similarly, when $s < 0.51$, the condition on Λ is also $\Theta(\Lambda/n) \sim 0.6$ for Google+ in overall structural information based DA. This is because when s is small, less common edges are shared by G^a and G^u. Consequently, it tends to involve all the anonymized users as seeds to achieve perfect de-anonymizability.

• For seed-based DA, when s is above some threshold value, $\Theta(\Lambda/n)$ decreases with the increases of s (less seed mappings are needed), e.g., when s is increased from 0.8 to 0.9, $\Theta(\Lambda/n)$ decreases from 0.47 to 0.3. For overall structural information based DA, when s is above some value, it is a.a.s. that a social network is $(1-\epsilon)$-de-anonymizable even without any seed mapping information, e.g., when $s \geq 0.51$, $\Theta(\Lambda/n) \sim 0$ for Google+. This is because: (i) when s increases, G^a and G^u are more structurally similar. Thus, G^a is more de-anonymizable in both seed and overall structural information based DA; and (ii) when the overall structural information is considered, the perfect DA scheme tends to induce the least edge difference when s is above some threshold value, i.e., a social network becomes $(1-\epsilon)$-de-anonymizable when s is large enough, which is also consistent with our quantification.

If we fix $s = 0.8$, the condition on Λ to make each social network $(1-\epsilon)$-de-anonymizable under different ϵ is shown in Fig.16. From Fig.16, we can see that:

• In seed-based DA, to make the social networks with low average degree $(1-\epsilon)$-de-anonymizable, it is necessary to identify $\Theta((1-\epsilon)n)$ seed mappings. For example, the social networks shown in Fig.16 (a)-(d) have $\overline{d} < 15$ and the condition on Λ to make them $(1-\epsilon)$-de-anonymizable is $\Theta(\Lambda/n) \sim 1-\epsilon$. The reason is that a low $\overline{d}$ implies less edges from anonymized users to seed users. Consequently, more seed mappings are necessary. On the other hand, if a social network has a large $\overline{d}$, e.g., most of the social networks in Fig.16 (e)-(h), less seed mappings are needed to be $(1-\epsilon)$-de-anonymizable in seed-based DA. For instance, when $\epsilon = 0.6$, to make Google+ 0.4-de-anonymizable, 22.52% users are needed to serve as seeds.

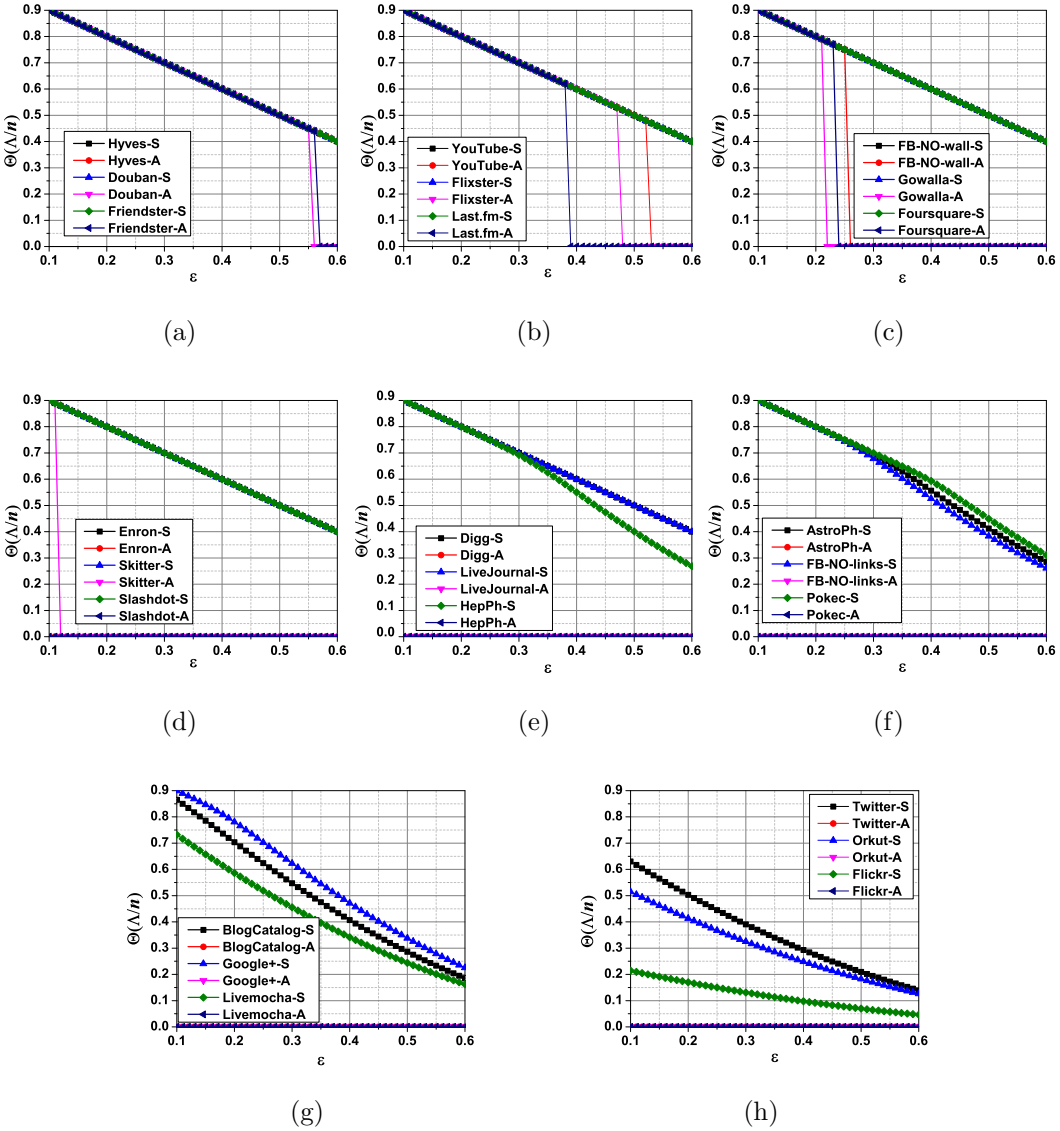


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 15: $(1-\epsilon)$-DA: Λ vs. s. Default setting: $\epsilon = 0.4$.



(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 16: $(1-\epsilon)$-DA: Λ vs. ϵ. Default setting: $s = 0.8$.

• In overall structural information based DA, if ϵ (the tolerated DA error) is above some threshold value, all the 24 social networks are $(1-\epsilon)$-de-anonymizable except for Hyves, which has a very low $\overline{d} = 3.96$. The reason is that when overall structural information (including seed mappings) is considered and $s = 0.8$, the correct DA induces the least edge difference with higher probability than in the seed-based DA, which is consistent with our quantification. Again, the results also confirmed that the overall structural information based DA is more effective.

Now, we evaluate the condition on Λ when the network size changes while other network properties are fixed. The results are shown in Fig.17. From Fig.17, we have the following observations.

• When n varies, the behavior of $\Theta(\Lambda/n)$ is similar to that when s varies. For the social networks with low $\overline{d}$, e.g., the social networks shown in Fig.17 (a)-(d), it is necessary to have $\Theta(\Lambda/n) \sim 1-\epsilon$ in seed-based DA. The reason is also similar to that presented in the earlier analysis. A small $\overline{d}$ implies less edges between anonymized users and seed users. Hence, it is necessary to have $\Theta(\Lambda/n) \sim 1-\epsilon$ to perfectly de-anonymize $(1-\epsilon)n$ users. On the other hand, when the network size is above some threshold value and continues to increase, less seed mappings are needed for social networks with high $\overline{d}$ (social networks in Fig.17 (e)-(h)) to be $(1-\epsilon)$-de-anonymizable. The reason is also similar to that presented in the earlier analysis.

• Again, the overall structural information based DA is more powerful, i.e., even without seed information, the structure itself can make a social network perfectly de-anonymizable. The quantification along with the evaluation results provides the foundation of the DA attack without seed information.

4.5.3.3 Evaluation on n

In this subsection, we evaluate the condition on n for the $(1-\epsilon)$-de-anonymizability of each social network. First, we examine $\Omega(n)$ under different settings of s. The results are shown in Fig.18, where $y \times n$ (y-axis) denotes the required network size is y times that of the original network size n. From Fig.18, we can see that:

• When s increases, the condition on n becomes loose for $(1-\epsilon)$-DA. For instance, when s increases from 0.5 to 0.9, the network size requirement to make Google+ $(1-\epsilon)$-de-anonymizable decreases from 2.85E8 to 3.19E7. This is because a large s implies more common edges between G^a and G^u followed by more structural similarity between them. Consequently, as shown in our quantification, it is still a.a.s. that the

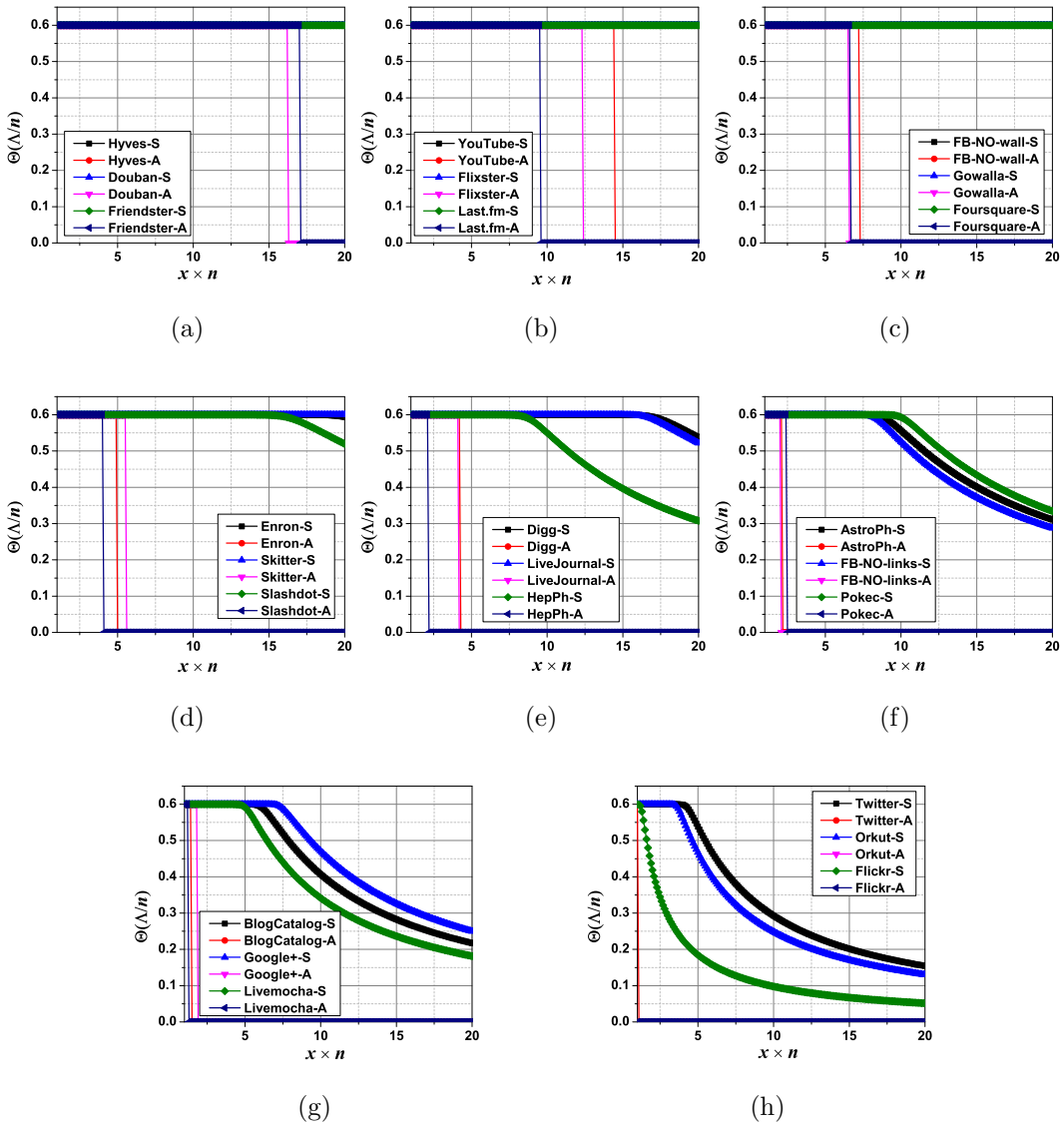


Figure 17: $(1-\epsilon)$-DA: Λ vs. n. Default setting: $s = 0.8$ and $\epsilon = 0.4$.

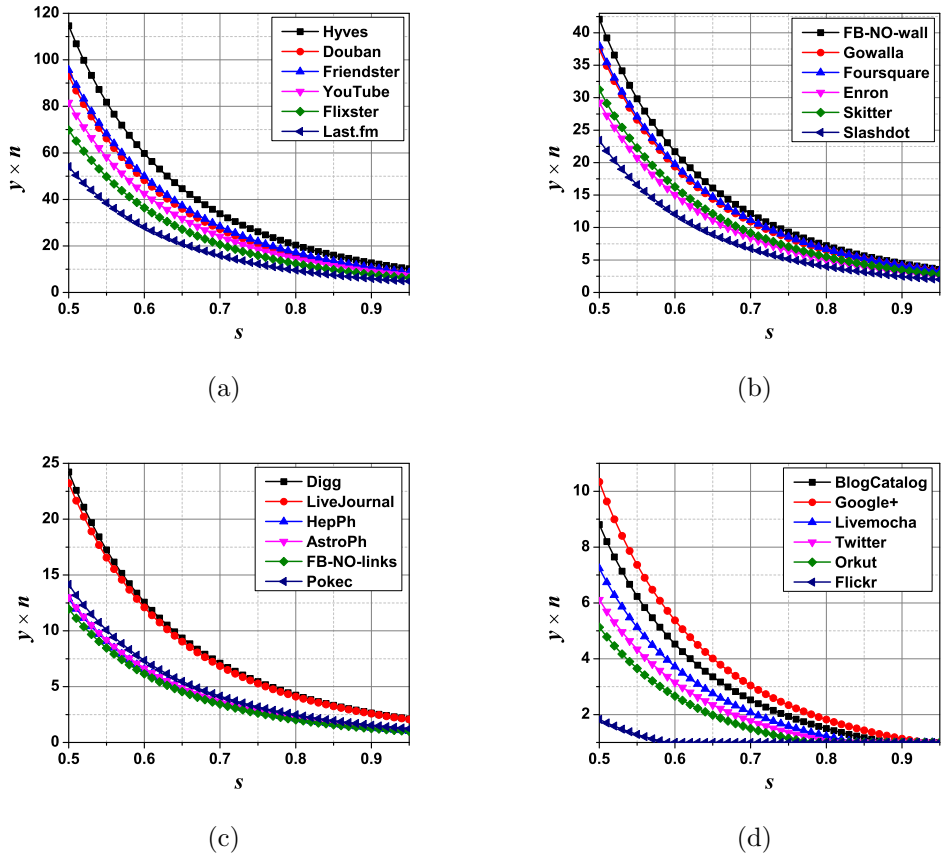


Figure 18: $(1-\epsilon)$-DA: n vs. s. Default setting: $\epsilon = 0.4$ and $\Lambda/n = 0.05$.

correct DA induces the least edge difference under such loose condition.

• Generally speaking, a high $\overline{d}$ implies loose condition requirement on n. From Table 9, the average degree range of the social networks shown in Fig.18 (a), (b), (c), and (d) are (3.96, 7.58), (8.01, 14.18), (15.32, 27.32), and (34.1, 146.56), respectively. It is evident that the social networks in Fig.18 (d) (e.g., Flickr, Orkut, Twitter.) have a looser requirement on network size than the social networks in Fig.18 (a) (e.g., Hyves, Douban, YouTube.). This is because a higher $\overline{d}$ implies richer structural information and more connections between anonymized users and seed users.

When the tolerated error ϵ is increased, the condition on network size for $(1-\epsilon)$-DA is shown in Fig.19. From Fig.19, we can see that:

• When ϵ increases, the condition on network size becomes loose. For instance,

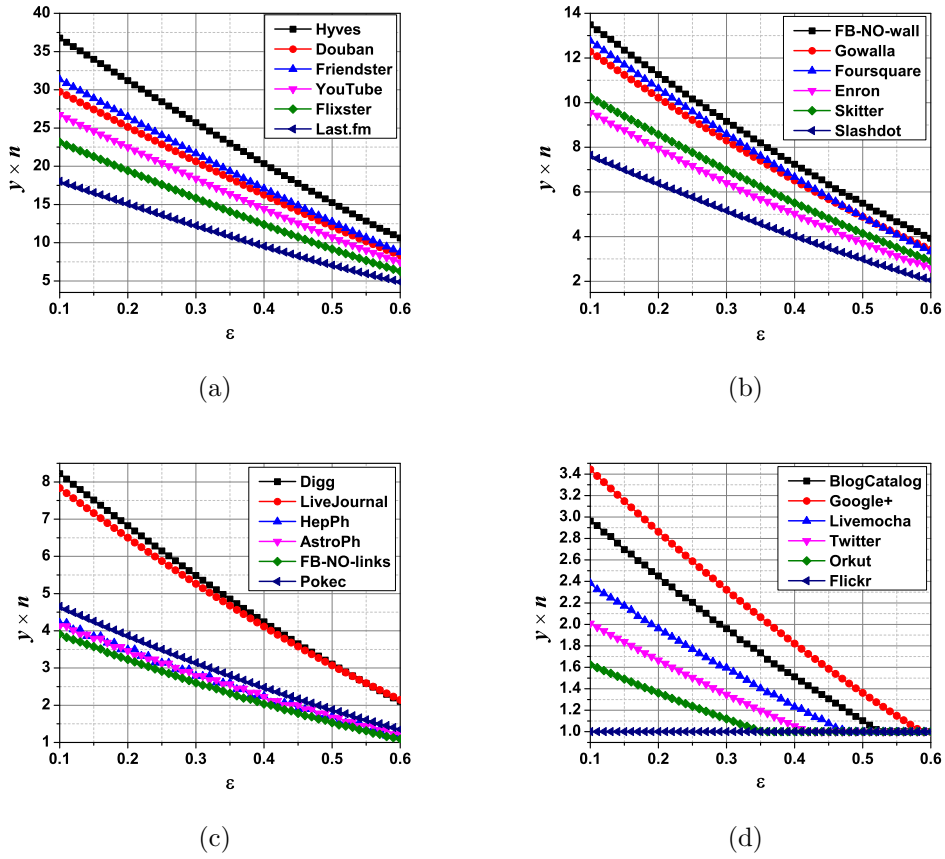


Figure 19: $(1-\epsilon)$-DA: n vs. ϵ. Default setting: $s = 0.8$ and $\Lambda/n = 0.05$.

when ϵ is increased from 0.1 to 0.4, the network size is changed from $2.01n$ to $1.05n$, where n is the size of the Twitter dataset, which implies Twitter is a.a.s. $(1-\epsilon)$-de-anonymizable at its current form when $\epsilon \geq 0.4$. The reason is evident since more tolerable error (i.e., a large ϵ) implies higher probability to be $(1-\epsilon)$-de-anonymizable, followed by a loose condition requirement on the network size.

- As in the evaluation of examining the network size while changing s, the social networks with higher $\overline{d}$ have a loose condition requirement on the network size, e.g., Flickr ($\overline{d} = 146.56$) and Orkut ($\overline{d} = 76.28$) have a looser network size requirement than that of Hyves ($\overline{d} = 3.96$) and Douban ($\overline{d} = 4.22$). The reason is the same as explained before. A higher $\overline{d}$ implies more structural similarity between G^a and G^u, followed by a loose condition on the network size as shown in our quantification.

Finally, we evaluate the condition on network size when various seed mappings are available. The results are shown in Fig.20 (note that, $y \geq 1$). From Fig.20, we have observations as follows.

• For seed-based DA, when $\Theta(\Lambda/n)$ increases, the required network size decreases fast. The reason is that if more seed mappings are available, more edges are expected to appear between the anonymized users and seed users, followed by a higher probability that the correct DA inducing the least edge difference. This is also consistent with our quantification.

• The number of available seed mappings has a limited impact on overall structural information based DA. This is because the overall structural information based DA scheme considers all the structural information simultaneously and seeks the DA which minimizes the overall edge difference. As long as $\Lambda/n < 0.5$, it is expected that the structural information carried by anonymized users dominates the DA process instead of the structural information carried by seed users (note that our seeds are randomly chosen).

• Similar to the results in Fig.18 and Fig.19, the social networks with high $\overline{d}$ requires a loose condition on network size given the same $\Theta(\Lambda/n)$ in general. The reason is also the same as that presented in the earlier analysis.

4.6 Chapter Summarization

In this chapter, we study the de-anonymizability of graph data based only on their structural information. First, we quantify the *perfect de-anonymizability* and $(1-\epsilon)$-*de-anonymizability* of graph data with seed information under the mathematical ER model. Subsequently, we extend our quantification to general scenarios, where a graph can follow an arbitrary model. To the best of our knowledge, this is the first comprehensive quantification study on the perfect and partial de-anonymizability of graph data with seed information under a general model. Third, based on our

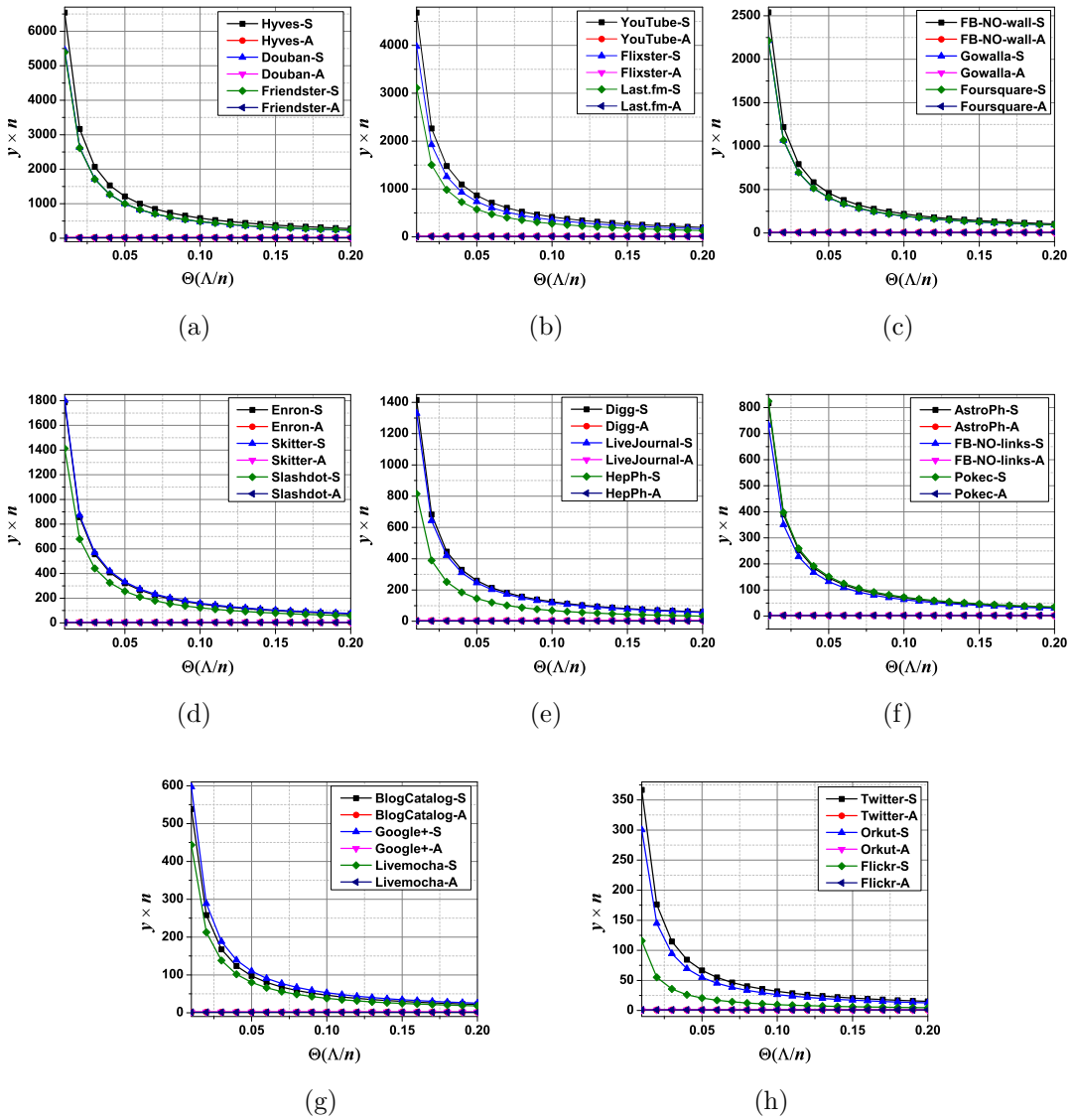


Figure 20: $(1-\epsilon)$-DA: n vs. Λ. Default setting: $s = 0.8$ and $\epsilon = 0.4$.

quantification, we conduct a large scale evaluation on the de-anonymizability of 24 various real world social networks. In the evaluation, we demonstrate the conditions on perfectly or partially de-anonymizing a social network, how many users of each social network can be successfully de-anonymized, etc. Furthermore, we show that, both theoretically and experimentally, the overall structural information based DA attack can be powerful, and even without any seed information available, a graph can also be perfectly or partially de-anonymizable. Our findings are expected to shed light on the future research in the structural data anonymization and DA areas, and help data owners evaluate their data vulnerability before data sharing/publishing.

CHAPTER V

DE-SAG: DE-ANONYMIZING SOCIAL ATTRIBUTE GRAPHS

5.1 Introduction

In most real scenarios of sharing/publishing graph data, in addition to sharing/publishing the graph structure, a lot of non-*Personal Identifiable Information* (non-PII), or *attribute information*, associated with graph users is also shared or published, e.g., gender, education, city, country, interests [15, 16]. Therefore, when studying anonymization and De-Anonymization (DA) techniques for graph data, the following question can be posed: *what are the impacts of the attribute information on the anonymity/de-anonymizability of graph data?* However, in existing graph data DA research [27, 64, 69, 104, 127, 151], only graph structure information is considered. Similarly, existing graph anonymity/de-anonymizability quantification research [61, 63, 69, 113, 151] only consider the graph structure, which gives an incomplete picture of the actual privacy vulnerability of graph data. To address the aforementioned open problem, we study the impact of attribute information (non-PII) on the privacy of graph data both theoretically and empirically. In this chapter, to distinguish between graph data with just graph structure and graph data with structure and attributes, we name the graph data with structure and attribute information *Structure-Attribute Graph* (SAG) data. Our main contributions in this chapter can be summarized as follows.

1. We conduct the first *attribute-based anonymity analysis* of SAG data under both preliminary and general data models. By careful quantification, we explicitly demonstrate the correlation between the achievable graph anonymity and the attribute information. Our theoretical results demonstrate that the attribute

information, even as non-PII, can also lead to significant anonymity loss of graph data. We also validate our analysis by both numerical evaluation and real world SAG data-based evaluation. The evaluation results further confirm our anonymity analysis. Our attribute-based anonymity analysis together with existing structure-based de-anonymizability quantifications provide data owners and researchers a more complete understanding of the privacy of graph data.

2. According to our attribute-based anonymity analysis, we propose a new DA attack on graph data, namely De-SAG, which takes into account both graph structure and attribute information to the best of our knowledge. Through extensive evaluations leveraging real world SAG data, we demonstrate that De-SAG can significantly enhance existing graph DA attacks. For instance, when de-anonymizing a Facebook dataset (4,039 users, 88,234 user-user links, 1,283 attributes, 37,257 user-attribute links), De-SAG has a 3.82 $\sim$ 10.1 times better DA performance than state-of-the-art structure-based DA attacks [63, 69].

The rest of this chapter is organized as follows. We provide the data model, preliminaries, and definitions in Section 5.2. The attribute-based anonymity analysis and evaluation are conducted in Section 5.3. Then, we propose and evaluate De-SAG in Section 5.4. The chapter is concluded in Section 5.5.

5.2 Data Model, Preliminaries, and Definitions

5.2.1 Data Model

Given a SAG, we model it as a graph $G = (V, E, A, W)$ as shown in Fig.21, where $V = \{i | i \text{ is a user}\}$ (the set of users), $E = \{l_{ij} | l_{ij} \text{ is a link between users } i \text{ and } j\}$ (the set of all the links among users), $A = \{i | i \text{ is an attribute}\}$ (the set of all the non-PII associated with the users in V), and $W = \{a_{ij} | i \in V, j \in A, a_{ij} \text{ is a link between user } i \text{ and attribute } j$, i.e., user i has attribute $j\}$ (the set of all the links between users and attributes). For $\forall i \in V$, we denote the attributes associated with i by $\mathcal{A}_i$, i.e.,

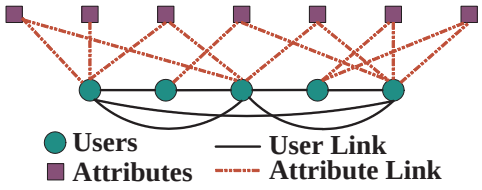


Figure 21: The SAG model.

$\mathcal{A}_i = \{j | j \in A, \exists a_{ij} \in W\}$. Furthermore, we define $n = |V|$ and $N = |A|$ to be the numbers of users and attributes, respectively.

5.2.2 De-anonymization

Given a raw SAG G, we assume that it will be anonymized before being shared/published. The anonymized G is denoted by $G' = (V', E', A', W')$ (we use an *apostrophe* to distinguish between the notations associated with G' from G when necessary). Note that, in G', although we cannot distinguish between the users in V' (we do not know the identities of the users in V'), we still know the attributes associated with each anonymized user since they are non-PII, e.g., in the published SAG data [15, 16, 49], the attributes (non-PII) associated with anonymized users are explicitly available. On the other hand, in reality, for $\forall i \in V$, it is also possible that $\mathcal{A}_i \neq \mathcal{A}'_i$ after the anonymization process, i.e., the anonymization scheme may add some new attributes to and/or remove some existing attributes from a user.

For the adversaries, as in existing DA attacks [63, 104, 108, 127], they try to de-anonymize G' leveraging some auxiliary graph denoted by $G'' = (V'', E'', A'', W'')$ (we use *double-apostrophe* to distinguish between the notations associated with G'' from G' and G when necessary), e.g., an adversary can leverage a Flickr graph to deanonymize a Twitter graph [104]. In reality, the auxiliary graphs can be obtained through multiple means, e.g., online crawling, data aggregation, data mining tasks, third-party information collection, public data sharing [63, 104].

Without loss of generality, we assume $V' = V'' = V$ (although we do not know

the users in V') and $A' = A'' = A$. Note that, as in [63, 113], this assumption does not limit the results of this chapter. When $V' \neq V''$ (respectively, $A' \neq A''$), the analysis in this chapter is valid on V'_{new} and V''_{new} (respectively, A'_{new} and A''_{new}) which are defined as $V'_{new} = V''_{new} = V' \cup V''$ (respectively, $A'_{new} = A''_{new} = A' \cup V''$); and the algorithm proposed in this chapter can still work directly.

According to G' and G'', a DA attack/scheme can mathematically be defined as a mapping from V' to V'' [61, 63, 104, 108], denoted by

$$\pi = V' \to V'' = \{(i, \pi(i) = j) | i \in V', j \in V''\}. \tag{150}$$

Then, for convenience of discussion, $\forall i \in V'$, a correct DA of i is denoted by mapping (i, i), i.e., the *identical mapping* corresponds to the correct DA.

5.2.3 Anonymity of G'

Entropy has been widely used to quantify the randomness/uncertainty of a process/system. Similarly, it can also be employed to measure the anonymity of G' given G'' and π [108]. Let π be an arbitrary DA scheme (mapping) from V' to V''. $\forall i \in V'$ and $\forall j \in V''$, let p^{π}_{ij} be the probability of the event that i *is mapped to* j *under* π. Then, for $\forall i \in V'$, we denote its *mapping distribution* under π as $\mathbf{P}^{\pi}_{i} =< p^{\pi}_{i1}, p^{\pi}_{i2}, \cdots, p^{\pi}_{in} >$. Hence, the *uncertainty* of i under π can be measured by the *entropy carried by the mapping distribution* $\mathbf{P}^{\pi}_{i}$, which is formally defined as

$$H^{\pi}(i) = -\sum_{j=1}^{n} p^{\pi}_{ij} \log p^{\pi}_{ij}. \tag{151}$$

Then, we define the *entropy/uncertainty of* G' under π as

$$H^{\pi}(G') = \frac{1}{n} \sum_{i=1}^{n} H^{\pi}(i), \tag{152}$$

which is the *average entropy* of all the users in G'.

Let $H_{\max}(i) = \max\{H^{\pi}(i)\}$ and $H_{\max}(G') = \{H^{\pi}(G')\}$, respectively. Evidently, for each $i \in V'$, $H^{\pi}(i)$ is maximized when i can be mapped to each user in V''

equiprobably. Hence, we have $H_{\max}(i) = \log n$. Similarly, we have $H_{\max}(G') = \log n$ when every user in G' achieves its maximum entropy. Then, based on $H^{\pi}(G')$ and $H_{\max}(G')$, we define the *anonymity* of G' under π as

$$\mathbb{A}(G') = \frac{H^{\pi}(G')}{H_{\max}(G')}. \tag{153}$$

From the definition, we have $\mathbb{A}(G') \in [0, 1]$, where a large value of $\mathbb{A}(G')$ implies a better anonymity of G'. Specifically, $\mathbb{A}(G') = 0$ implies all the users in G' can be successfully de-anonymized under π while $\mathbb{A}(G') = 1$ implies that G' achieves the perfect anonymity.

5.3 Anonymity Analysis: From the Attribute Perspective

As we discussed in Sections 5.1, the structure-based de-anonymizability analysis for graph data has been studied in [61, 63, 69, 113, 151]. However, understanding the impacts of attributes on the anonymity/de-anonymizability of graph data is still an open problem. Furthermore, no existing DA scheme employs both the graph structure and the associated attributes to de-anonymize graph data. In this section, we address the first open problem by measuring the impacts of attributes on SAG data's anonymity. To be mathematically tractable, we conduct the analysis under a preliminary model first. Then, we generalize the analysis to the more complicated practical scenarios.

5.3.1 Preliminary Analysis

First, we conduct *attribute-based anonymity analysis* for SAGs under a random *Attribute Attachment* (A^2) model: given a SAG G, we assume that *for* $\forall i \in V$ *and* $\forall j \in A$, *the existing probability of link* a_{ij} *is* p, i.e., $\Pr(a_{ij} \in W | \forall i \in V, \forall j \in A) = p$. Furthermore, we assume W' and W'' are random subsets of W: for each user-attribute link in W, it appeares in W' and W'' with positive probabilities p' and p'', respectively, i.e., $\Pr(a_{ij} \in W' | a_{ij} \in W) = p'$ and $\Pr(a_{ij} \in W'' | a_{ij} \in W) = p''$.

To facilitate our analysis, we introduce the concept of *Attribute Difference* (AD) between the users in V' and V''. $\forall i \in V'$ and $\forall j \in V''$, their AD is defined as

$$D_{ij} = (\mathcal{A}'_i \cup \mathcal{A}''_j) \setminus (\mathcal{A}'_i \cup \mathcal{A}''_j). \tag{154}$$

Let $\alpha = pp'(1-p'')+pp''(1-p')$ and $\beta = pp'(1-pp'')+pp''(1-pp')$. Furthermore, let $\vartheta = \frac{(\beta-\alpha)^2}{8(\beta+\alpha)}$. Then, we have the following theorem which quantifies the *attribute-based anonymity loss* of G'.

Theorem 20. *Let t be a natural number and $t \in [1, n-1]$. Then, (i) if $\vartheta \geq \frac{2\ln n + t\ln(n-1) - \ln t! + 1}{Nt}$, $\mathbb{A}(G') = \frac{\log t}{\log n}$; and (ii) if $\vartheta \geq \frac{3\ln n + t\ln(n-1) - \ln t! + 1}{Nt}$, $\mathbb{A}(G') = 0$, i.e., G' lost all of the anonymity.*

Proof Sketch: (i) To prove this conclusion, we first analyze the entropy of $\forall i \in V'$. Suppose i is de-anonymized to $j \in V''$ under some DA scheme π (we will discuss how to determine π later), i.e., $\pi(i) = j$. Then, we analyzed the AD caused by mapping (i, j). On one hand, if $j = i$, an AD will be induced if i has one attribute in exactly one of V' and V''. It follows that the AD corresponding to mapping $(i, j = i)$ is $D_{ij} = D_{ii} \sim \mathbf{B}(N, \alpha)$, where $\mathbf{B}(N, \alpha)$ is a *binomial variable* with parameters N and α. On the other hand, if $j \neq i$, the AD corresponding to mapping (i, j) is $D_{ij} \sim \mathbf{B}(N, \beta)$. Clearly, $\beta > \alpha$.

Let $\mathbf{E}$ be the event that $\exists j \neq i$ *such that* $D_{ii} \geq D_{ij}$. Then, according to Lemma 1, we have

$$\Pr(\mathbf{E}) \leq 2\exp(-\frac{(N\beta - N\alpha)^2}{8(N\beta + N\alpha)}) = 2\exp(-N\vartheta). \tag{155}$$

Furthermore, the possible number of such events can be counted by t. Let $\mathbf{E}_t$ be the

event that $\mathbf{E}$ *happens* t *times*. Then, we have

$$\Pr(\mathbf{E}_t) = C(n-1,t) \cdot \Pr(\mathbf{E})^t \cdot (1-\Pr(\mathbf{E}))^{n-t} \tag{156}$$

$$\leq C(n-1,t) \cdot \Pr(\mathbf{E})^t \tag{157}$$

$$\leq \frac{(n-1)^t}{t!} \cdot 2\exp(-N\vartheta t) \tag{158}$$

$$= \exp(t\ln(n-1) - \ln t!) \cdot 2\exp(-N\vartheta t) \tag{159}$$

$$= 2\exp(t\ln(n-1) - \ln t! - N\vartheta t) \tag{160}$$

$$\leq 2\exp(-2\ln n - 1) \tag{161}$$

$$\leq \frac{1}{n^2}. \tag{162}$$

According to the Borel-Cantelli Lemma, we have $\Pr(\mathbf{E}_t) \to 0$ as $n \to \infty$. Therefore, when $\vartheta \geq \frac{2\ln n + t\ln(n-1) - \ln t! + 1}{Nt}$, with probability 1, $\mathbf{E}$ *happens less than* t *times*.

Algorithm 2: An implementation of π

for $i \in V'$ **do**
  sorting the users in V'' in the increasing order of D_{ij} for $j \in V''$ and the sorted sequence is denoted as $< j_1, j_2, \cdots, j_n >$;
  mapping i to j_k $(1 \leq k \leq t)$ with probability $\frac{1}{t}$;

Based on the analysis, we define a simple de-aonymization scheme π as shown in Algorithm 2. From Algorithm 2, we have $\mathbf{P}_i^\pi =< p_{j_1}^\pi, p_{j_2}^\pi, \cdots, p_{j_t}^\pi, p_{j_{t+1}}^\pi, \cdots, p_{j_n}^\pi > =< \frac{1}{t}, \frac{1}{t}, \cdots, \frac{1}{t}, 0, \cdots, 0 >$. Furthermore, considering that $\Pr(\mathbf{E}_t) \to 0$, we conclude that π can successfully de-anonymize any user in V' with probability $\frac{1}{t}$. Then, $\forall i \in V'$, we have $H^\pi(i) = \log t$. It follows that $H^\pi(G') = \log t$ and thus $\mathbb{A}(G') = \frac{H^\pi(G')}{H_{\max}(G')} = \frac{\log t}{\log n}$.

(ii) Now, we prove the second conclusion. Let $\mathbf{E}_{all}$ be the event that *there exists some* t *such that* $\mathbf{E}_t$ *happens*. Then, $\Pr(\mathbf{E}_{all}) = \bigcup_{t=1}^{n} \Pr(\mathbf{E}_t)$. Based on the *Boole's*

inequality, we have

$$\Pr(\mathbf{E}_{all}) = \bigcup_{t=1}^{n} \Pr(\mathbf{E}_t) \tag{163}$$

$$\leq \sum_{t=1}^{n-1} \Pr(\mathbf{E}_t) \tag{164}$$

$$\leq \sum_{t=1}^{n-1} 2\exp(t\ln(n-1) - \ln t! - N\vartheta t) \tag{165}$$

$$= \sum_{t=1}^{n-1} 2\exp(-3\ln n - 1) \tag{166}$$

$$\leq \frac{1}{n^2}. \tag{167}$$

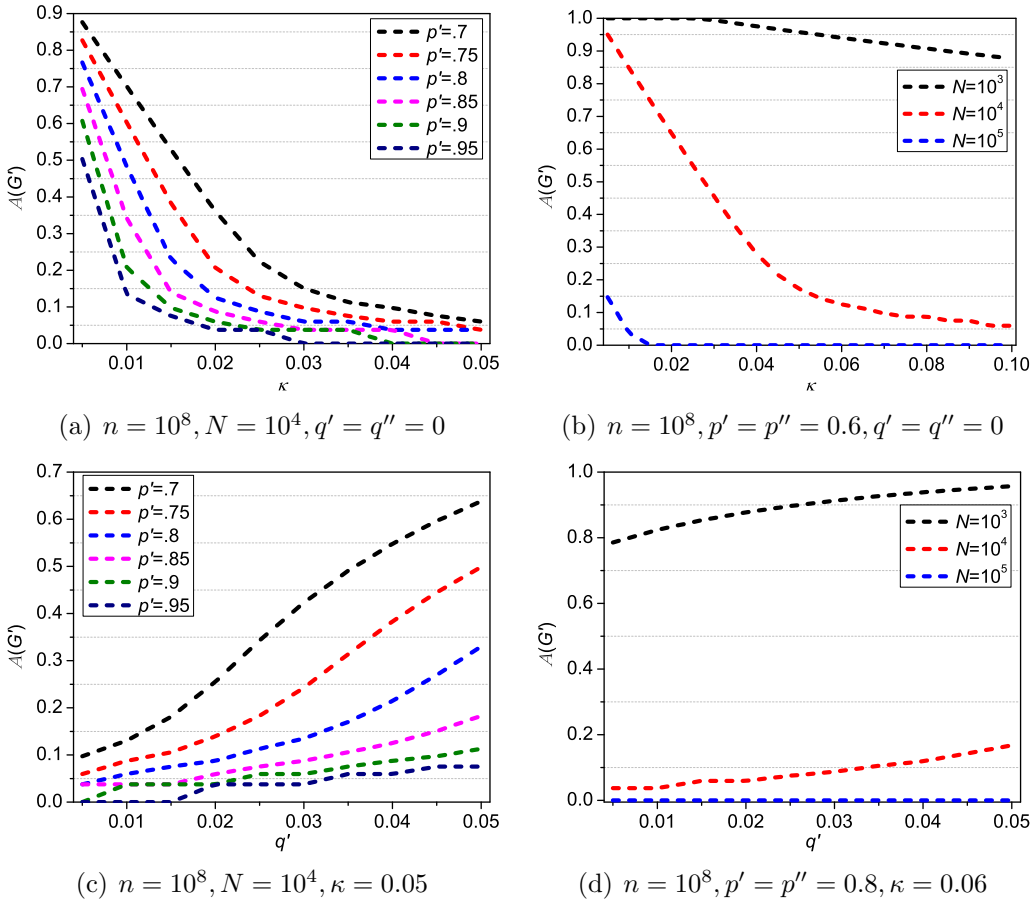
According to the Borel-Cantelli Lemma, we have $\Pr(\mathbf{E}_{all}) \to 0$ as $n \to \infty$, i.e., when $\vartheta \geq \frac{3\ln n + t\ln(n-1) - \ln t! + 1}{Nt}$, $\nexists t$ such that $\mathbf{E}_t$ happens. This further implies that with probability 1, $\forall i \in V'$ and $\forall j \in V''$, if $j \neq i$, $\Pr(D_{ii} < D_{ij}) \to 1$ as $n \to \infty$.

Algorithm 3: Another implementation of π

```
for i ∈ V' do
    mapping i to j ∈ V'' such that j = arg min_j {D_ij | j ∈ V''};
```

Based on our analysis, we give another simple implementation of π as shown in Algorithm 3. Under π, each user in V' can be successfully de-anonymized with probability 1 as $n \to \infty$. Therefore, $H^{\pi}(i) = 0$ for $\forall i \in V'$. It follows $\mathbb{A}(G') = 0$, i.e., all the users can be successfully de-anonymized by π with probability 1. □

In Theorem 20, we analyzed the impacts of attributes (non-PII) on the anonymity/de-anonymizability of SAG data under the A^2 model. Based on our analysis, the attributes may also significantly reduce the anonymity of SAG data, which is similar to the graph structure (as shown in [61, 63, 113]). To make our analysis more practical, we extend it to general scenarios in the following subsection.



(a) $n = 10^8, N = 10^4, q' = q'' = 0$ (b) $n = 10^8, p' = p'' = 0.6, q' = q'' = 0$

(c) $n = 10^8, N = 10^4, \kappa = 0.05$ (d) $n = 10^8, p' = p'' = 0.8, \kappa = 0.06$

Figure 22: Numerical evaluation of $\mathbb{A}(G')$.

5.3.2 Extension: Practical Scenarios

In the analysis under the A^2 model, W' and W'' are two random subsets of W, which implies that $\forall i \in V$, $\mathcal{A}'_i$ and $\mathcal{A}''_i$ are two random subsets of $\mathcal{A}_i$. However, in reality, it is possible that some attributes in $\mathcal{A}_i$ may not appear in $\mathcal{A}'_i/\mathcal{A}''_i$ or some attributes in $A \setminus \mathcal{A}_i$ may appear in $\mathcal{A}'_i/\mathcal{A}''_i$. Therefore, in this subsection, we conduct the *attribute-based anonymity analysis* for SAG data under a more general model.

Under the general model, $\forall a_{ij} \in W$, it is appeared in W' and W'' with probabilities p' and p'' respectively, i.e., $\Pr(a_{ij} \in W' | a_{ij} \in W) = p'$ and $\Pr(a_{ij} \in W'' | a_{ij} \in W) = p''$. Furthermore, $\forall a_{ij} \notin W$, it is appeared in W' and W'' with probabilities q' and q'' respectively, $\Pr(a_{ij} \in W' | a_{ij} \notin W) = q'$ and $\Pr(a_{ij} \in W'' | a_{ij} \notin W) = q''$. Let $W_U = \{a_{ij} | i \in V, j \in A\}$ be the *universal set* of all the possible user-attribute links.

Then, $\forall a_{ij} \in W_U$, we have $\Pr(a_{ij} \in W | a_{ij} \in W_U) \underset{statistically}{\rightarrow} \frac{|W|}{|W_U|}$. Let $\kappa = \frac{|W|}{|W_U|}$ and define $\zeta = \kappa(p'(1-p'') + p''(1-p')) + (1-\kappa)(q'(1-q'') + q''(1-q'))$, $\delta = (\kappa p' + (1-\kappa)q')(\kappa(1-p'') + (1-\kappa)(1-q'')) + (\kappa(1-p') + (1-\kappa)(1-q'))(\kappa p'' + (1-\kappa)q'')$, and $\varpi = \frac{(\delta-\zeta)^2}{8(\delta+\zeta)}$. Then, we have the following theorem to quantify the impacts of attributes on the achievable anonymity of G'.

Theorem 21. *Let t be a natural number and $t \in [1, n-1]$. Then, (i) if $\delta > \zeta$ and $\varpi \geq \frac{2\ln n + t\ln(n-1) - \ln t! + 1}{Nt}$, $\mathbb{A}(G') = \frac{\log t}{\log n}$; and (ii) if $\delta > \zeta$ and $\varpi \geq \frac{3\ln n + t\ln(n-1) - \ln t! + 1}{Nt}$, $\mathbb{A}(G') = 0$.*

Proof: This theorem can be proven using similar techniques as in Theorem 20. □

In Theorem 21, we show the achievable anonymity of G' under a general statistical model. From Theorem 21, the condition on ϖ is similar to that of ϑ in Theorem 20. In addition, Theorem 21 has one more constraint $\delta > \zeta$, which actually comes from the fact that *for $i \in V$, the attributes that do not appeare in $\mathcal{A}_i$ may appear in $\mathcal{A}'_i$ and/or $\mathcal{A}''_i$*. Similar to Theorem 20, Theorem 21 also implies that the attributes associated with users (non-PII) may have significant impacts on the anonymity of SAG data.

5.3.3 Evaluation

In this subsection, we evaluate our *attribute-based anonymity analysis* both numerically and via experiments that leverage real world SAG datasets. Since there exists randomness in our evaluations, we repeat each group of evaluations 100 times. The results are the average of these 100 evaluations.

5.3.3.1 Numerical Evaluation

Since the analysis under the A^2 model can be viewed as a special case of that under the general model, our numerical evaluation follows the anonymity analysis under the general model, i.e., Theorem 21. Furthermore, to simplify the evaluation process, we set $p' = p''$ and $q' = q''$. Note that this setting does not limit our evaluation, and it can be removed directly by considering more scenarios.

Table 10: Data statistics.

	n	m	N	M	κ
GP1	4,693,129	47,130,325	991,545	3,644,103	7.83E-07
GP2	17,091,929	271,915,755	3,108,141	14,693,125	2.77E-07
GP3	26,244,659	410,445,770	4,147,389	19,344,382	1.78E-07
GP4	28,942,911	462,994,069	4,443,631	20,592,962	1.60E-07
GP5	107,614	13,673,453	19,044	387,261	1.89E-04
Facebook	4,039	88,234	1,283	37,257	7.19E-03
Twitter	81,306	1,768,149	216,839	1,245,234	7.06E-05

In our evaluation, we first randomly generate a SAG G with the specified n, N, and κ. Subsequently, we generate G' and G'' from G according to p', p'', q', and q''. Finally, we evaluate $\mathbb{A}(G')$ based on Theorem 21. The detailed parameter settings are specified in each group of evaluations.

We show the evaluation results in Fig.22. We analyze the results as follows.

1. From Fig.22 (a), with the increase of κ, $\mathbb{A}(G')$ decreases under different p'. This is because a larger κ implies more attribute information is associated with each user, statistically. Therefore, different users are more distinguishable with respect to attributes, i.e., with a higher probability $D_{ii} \leq D_{ij}$. Furthermore, given κ, better anonymity is achieved when p' is smaller, e.g., given $\kappa = 0.02$, $\mathbb{A}(G') = 0.361$ when $p' = 0.7$ while $\mathbb{A}(G') = 0.038$ when $p' = 0.95$. This is because a larger p' implies more attributes can be preserved in G' and G'' (since $p'' = p'$), and thus it is more likely that $D_{ii} < D_{ij}$, i.e., a large p' implies more anonymity loss, which is consistent with our theoretical analysis.

2. From Fig.22 (b), given n, p', p'', q', and q'', $\mathbb{A}(G')$ decreases when κ increases under different N. The reason is the same as in Fig.22 (a): a larger κ implies a higher probability of $D_{ii} < D_{ij}$, i.e., more anonymity loss. In addition, given κ, a larger N also implies more anonymity loss. For instance, given $\kappa = 0.065$,

$\mathbb{A}(G') = 0.932$ when $N = 10^3$ while $\mathbb{A}(G') = 0.113$ when $N = 10^4$. This is because when κ is fixed, a larger N also implies richer attributes associated with each user and thus $D_{ii} < D_{ij}$ happens with a higher probability, i.e., $\mathbb{A}(G')$ decreases.

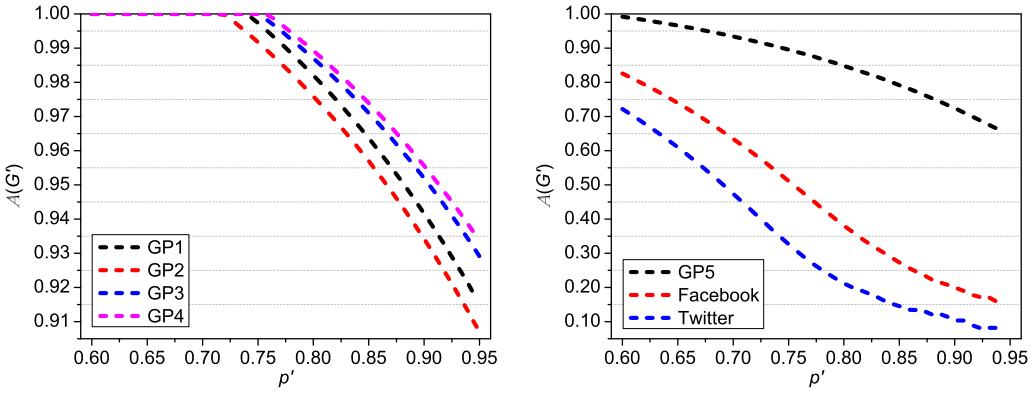
3. Fig.22 (c) shows the impacts of q' (q'') on $\mathbb{A}(G')$. From Fig.22 (c), when q' increases, $\mathbb{A}(G')$ increases under different p'. This is because q' indicates the percentage of fake user-attribute relationships being added to G' and G'' ($q'' = q'$). A larger q' implies more link noise has been added to W' and W'' and thus a lower probability of $D_{ii} < D_{ij}$ has been induced, followed by the increase of $\mathbb{A}(G')$. Furthermore, given q', a smaller p' implies more anonymity can be achieved. The reason is the same as analyzed before: a smaller p' implies less common attributes are shared between G' and G''. Hence, better anonymity can be achieved by G'.

4. In Fig.22 (d), we examine the impacts of q' and N on $\mathbb{A}(G')$. Again, when q' increases, $\mathbb{A}(G')$ also increases under different N. Additionally, given q', a larger N implies more anonymity loss. The reason is also the same as before. A larger N implies more attribute information is available for each user followed by less achievable anonymity according to our theoretical analysis.

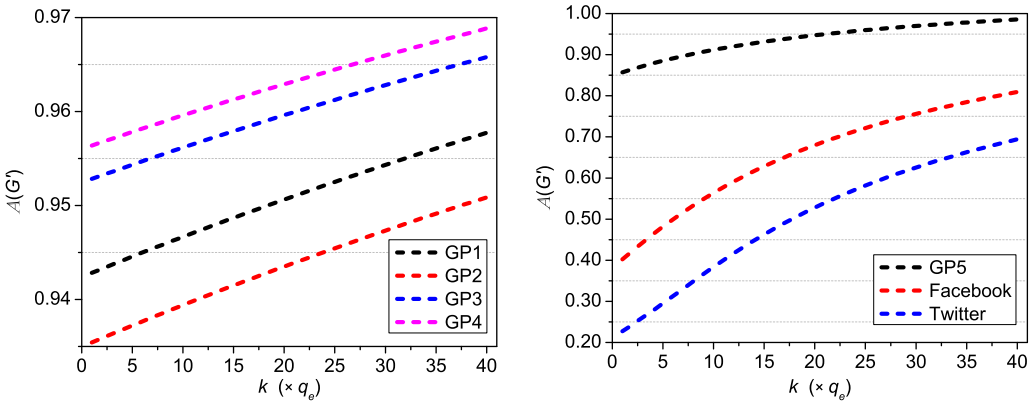
5.3.3.2 Real World Data-based Evaluation

Now, we evaluate our attribute-based anonymity analysis leveraging real world SAG datasets.

Datasets. The employed SAG datasets include five Google Plus (GP) datasets, denoted by GPk ($1 \leq k \leq 5$) respectively [15, 16, 49], one Facebook dataset [16], and one Twitter dataset [16] as shown in Table 10, where $n = |V|$ (the number of users), $m = |E|$ (the number of user-user links), $N = |A|$ (the number of attributes), $M = |W|$ (the number of user-attribute links), and $\kappa = \frac{M}{n \cdot N}$ (the *connectivity* between users



(a) $q' = q'' = 0$ (b) $q' = q'' = 0$



(c) $p' = p'' = 0.8$ (d) $p' = p'' = 0.8$

Figure 23: Evaluation of $\mathbb{A}(G')$ leveraging on real data.

and attributes). All the SAG datasets include both the graph structure information and the attribute information (non-PII) associated with users. We introduce the datasets as follows.

- GP is a social networking service launched in June 2011. It is designed to be a place to connect with friends and family. GP1, GP2, GP3, and GP4 are four GP datasets crawled in July 2011, August 2011, September 2011, and October 2011, respectively [15, 49]. In addition to the social relationship information, there is also attribute information in the four GP datasets, e.g., gender, affiliation information, education, city. They are available under application. GP5 is another GP dataset which is publicly available at [16]. The attribute information in GP5 includes education, hometown, language, etc.

- Facebook is one of the most popular social networking services in the world. It is designed as a social utility that connects people with friends and others who work, study, and live around them. The employed Facebook dataset is publicly available at [16]. The attribute information in Facebook includes birthday, education, location, employer, etc.

- Twitter is also a popular online social networking service that enables users to send and read short messages named *tweets*. The employed Twitter dataset is publicly available at [16]. The attribute information in the Twitter dataset includes interests, cities, sports, websites, etc.

Results and Analysis. When conducting real world SAG data based evaluation, we first generate G' and G'' from each dataset according to the specified p', p'', q', and q''. Then, we evaluate the anonymity of G' following our analysis in Theorem 21. The evaluation results are shown in Fig.23, where the parameters are specified in each group of simulations. We analyze Fig.23 as follows:

1. In Fig.23 (a) and (b), we show the impacts of p' on the achievable anonymity of each dataset, from which we can see that with the increase of p', the anonymity of each dataset decreases. For instance, when p' is increased from 0.6 to 0.8, $\mathbb{A}$(Facebook) is decreased from 0.826 to 0.381. The reason is similar to that in the numerical evaluation. A larger p' implies more attribute information is preserved in G' and G''. This further implies that the probability of $D_{ii} < D_{ij}$ increases, followed by the decrease of the anonymity. From Fig.23 (a) and (b), we can also see that the anonymity of GP1, GP2, GP3, and GP4 are higher than that of GP5, Facebook, and Twitter, e.g., when $p' = 0.85$, $\mathbb{A}(\text{GP1}) = 0.964$ while $\mathbb{A}(\text{Twitter}) = 0.145$. The main reason is that the connectivity of GPk $(1 \leq k \leq 4)$ is much smaller than that of GP5, Facebook, and Twitter. Therefore, GPk $(1 \leq k \leq 4)$ can achieve better anonymity, which is consistent

with our analysis.

2. Fig.23 (c) and (d) show the impacts of q', which is defined as $q' = k \cdot q_e = k \cdot \frac{(1-p')M}{|nN-M|}$ $(1 \leq k \leq 40)$ [1], on the anonymity of the seven datasets. From Fig.23 (c) and (d), the anonymity of each dataset increases with the increase of q', e.g., when q' is increased from $5q_e$ to $35q_e$, the anonymity of Facebook is increased from 0.482 to 0.785. The reason is that when q' increases, more fake user-attribute links (noise links) will be added to G' and G''. Then, the probability of $D_{ii} < D_{ij}$ is decreased, followed by the increase of the anonymity, which is consistent with our analysis.

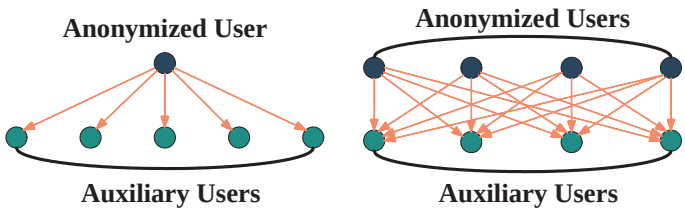
5.3.4 Discussion

From Section 5.1, structure-based de-anonymizability analysis for graph data has been conducted in [61, 63, 69, 113, 151]. In this chapter, for the first time, we study the impacts of attributes (non-PII) on the anonymity of graph data both theoretically and experimentally. Based on our analysis and evaluation results, we find that the attribute information associated with users may significantly reduce graph data anonymity. Therefore, our study together with existing structure-based de-anonymiability quantification research provides a much more complete understanding on the anonymity of graph data.

5.4 De-anonymization

In this section, we present a new DA framework, namely *De-SAG*, which considers both the graph structure and the attributes associated with users. Since *graph structure-based DA* has been well studied, we design De-SAG on top of existing structure-based DA attacks. Therefore, De-SAG can be considered as an enhanced

[1] Here, we set q' in terms of p'. This is because we do not want to add too many fake user-attribute links in G' and G'' compared to the number of removed real user-attribute links. Otherwise, the data utility of G' for data mining tasks might be ruined.



(a) (b)

Figure 24: User-based DA and set-based DA.

version of existing DA attacks. To facilitate our design, we define a new notation $V''_{i,t}$ for $i \in V'$, which denotes the *t-most similar users of i in V″ with respect to attributes*. Evidently, $V''_{i,t}$ can be obtained using the same technique as in Algorithm 2 (i.e., let $V''_{i,t} = \{j_k | k = 1, 2, \cdots, t\}$). Let $D_{\max} = \max\{D_{ij} | i \in V', j \in V''\}$ be the maximum AD between any user in V' and any user in V''. Then, we define the *attribute similarity* of $i \in V'$ and $j \in V''$ as $1 - \frac{D_{ij}}{D_{\max}}$.

5.4.1 De-SAG

With respect to the *DA process*, existing structure-based DA attacks can be classified as *user-based DA schemes*, e.g., [69, 104, 108, 151], and *set-based DA schemes*, e.g., [63, 64, 127] (the detailed explanations are given later). Since De-SAG is proposed on top of existing DA attacks, we present two implementations of the De-SAG framework based on the two classes of DA schemes.

5.4.1.1 User-based De-SAG

In user-based DA schemes [69, 104, 108, 151], as shown in Fig.24 (a), during each DA iteration, one anonymized user i in V' is selected based on some criteria (e.g., having the maximum degree, having the most number neighbors being de-anonymized, having the most number of seed neighbors). Then, i is mapped (de-anonymized) to some user in V'' according to the proposed DA technique and the next DA iteration is started.

To enhance existing user-based DA attacks, we present an implementation of the

Algorithm 4: User-based De-SAG

```
while V' ≠ ∅ do
    select i from V' as the user for DA according to the criteria in [69, 104, 108, 151];
    map i to some user j in V''_{i,t} according to the enhanced structure-based DA technique in [69, 104, 108, 151], i.e., taking the attribute similarity as an extra mapping feature;
    V' = V' \ {i};
    V'' = V'' \ {j};
```

user-based version of De-SAG as shown in Algorithm 4. From Algorithm 4, De-SAG basically follows the same process of existing user-based DA attacks. The primary improvements are (i) when de-anonymizing $i \in V'$, instead of considering all the users in V'' as candidates, we select the t-most-similar users of i with respect to attributes from V'' as candidate mappings. Here, t is a pre-defined parameter which controls the trade-off between DA accuracy and efficiency (a theoretically optimal t can be approximately estimated based on our anonymity analysis in Section 5.3[2]); and (ii) i is de-anonymized to one of the t-most similar users according to an *enhanced* version of existing structure-based DA attacks. In existing attacks, i is mapped to some user j in V'' according to the similarity of i and j's structural features, e.g., degree, betweenness centrality, closeness centrality [69, 104, 108, 151]. In the enhanced version of existing attacks, De-SAG takes the attribute similarity as an extra mapping feature.

Let $O(T)$ and $O(S)$ be the time and space complexities of the enhanced user-based DA scheme in Algorithm 4, respectively. Then, the time complexity of De-SAG in Algorithm 4 is upper bounded by $O(n^2 \log n + T)$ since the candidate set size is reduced (the $O(n^2 \log n)$ time complexity is used to compute $V''_{i,t}$). The actual

[2]To estimate the theoretically optimal t, we first specify a temporary mapping from V' to V''. For instance, we can simply mapping V' to V'' according to the users' degree sequence: sorting the users in V' and V'' according to the degree non-increasing order and denoting the obtained user sequences as $< i_1, i_2, \cdots, i_n >$ and $< j_1, j_2, \cdots, j_n >$, respectively; and mapping i_k to j_k for $1 \leq k \leq n$. Second, we estimate G as $G = (V = V' = V'', E = E' \cup E'', A, W = W' \cup W'')$. Third, we estimate p, p', and p'' based on G, G', and G''. Fourth, we can estimate ϑ in terms of p, p', and p''. Finally, we estimate t as $t = \arg\min_t \vartheta \geq \frac{2 \ln n + t \ln(n-1) - \ln t! + 1}{Nt}$.

time complexity of De-SAG depends on the particular enhanced structure-based DA attack. The space complexity of De-SAG is also $O(S)$, i.e., De-SAG does not increase the space complexity of the enhanced scheme.

5.4.1.2 Set-based De-SAG

In set-based DA schemes [63, 64, 127], as shown in Fig.24 (b), during each DA iteration, a subset of un-de-anonymized users $\widetilde{V}'$ is selected from V' and a subset of auxiliary users $\widetilde{V}''$ is selected from V'', respectively. Subsequently, a *complete weighted bipartite graph* $\widetilde{G} = (\widetilde{V}, \widetilde{E})$ with $\widetilde{V} = \widetilde{V}' \cup \widetilde{V}''$ and $\widetilde{E} = \{l_{ij} | i \in \widetilde{V}', j \in \widetilde{V}''\}$ is constructed, where the weight of each link l_{ij}, denoted by $w(l_{ij})$, is determined according to the proposed DA techniques (usually, the weight of link l_{ij} measures how structurally similar i and j are and a larger weight means they are more structurally similar) [63, 64, 127]. After constructing $\widetilde{G}$, the DA problem reduces to a *Maximum Weighted Bipartite graph Matching problem* (MWBM) on $\widetilde{G}$. Finally, by addressing the MWBM problem on $\widetilde{G}$ (e.g., using the Hungarian algorithm), a mapping from $\widetilde{V}'$ to $\widetilde{V}''$ can be determined and the next DA iteration is started.

Algorithm 5: Set-based De-SAG

while $V' \neq \emptyset$ **do**
  determine $\widetilde{V}' \subseteq V'$ according to the criteria in [63, 64, 127];
  determine $\widetilde{V}'' \subseteq V''$ according to the criteria in [63, 64, 127];
  for $i \in \widetilde{V}'$ **do**
    $\widetilde{V}''_{i,t} \leftarrow V''_{i,t} \cap \widetilde{V}''$;
  construct a bipartite graph $\widetilde{G} = (\widetilde{V}' \cup \widetilde{V}'', \widetilde{E})$, where $\widetilde{E} = \{l_{ij} | i \in \widetilde{V}', j \in \widetilde{V}''_{i,t}\}$;
  for $l_{ij} \in \widetilde{E}$ **do**
    determine $w(l_{ij})$ according to the technique in [63, 64, 127];
    $w_a \leftarrow 1 - \frac{D_{ij}}{D_{\max}}$;
    $w(l_{ij}) \leftarrow c \cdot w(l_{ij}) + (1 - c) \cdot w_a$;
  de-anonymize $\widetilde{V}'$ based on $\widetilde{G}$ using the technique in [63, 64, 127];
  subtract the de-anonymized users from V' and V'', respectively;

To enhance the set-based DA schemes [63, 64, 127], we present a set-based implementation of De-SAG as shown in Algorithm 5, where w_a denotes the *attribute similarity* of $i \in V'$ and $j \in \widetilde{V}''_{i,t}$, and $c \in [0, 1]$ is a pre-defined constant value. From Algorithm 5, De-SAG basically follows a similar process as existing set-based DA attacks. During each DA iteration, it improves existing schemes by further leveraging the attribute information. Specifically, De-SAG enhances existing set-based DA attacks in two perspectives. First, instead of constructing a *complete* bipartite graph, it reduces the number of links in $\widetilde{G}$ by setting $\widetilde{E} = \{l_{ij} | i \in \widetilde{V}', j \in \widetilde{V}''_{i,t}\}$. Second, it resets the weight associated with each link by taking account of the attribute similarity of two users (using $w(l_{ij}) \leftarrow c \cdot w(l_{ij}) + (1 - c) \cdot w_a$). Leveraging the two enhancements, (i) the computational complexity of existing set-based DA schemes can be reduced (since the mapping problem now is addressed on a non-complete bipartite graph); and (ii) the performance of existing set-based DA attacks can be improved (since the attribute similarity is used to enhance the DA process).

Let $O(T)$ and $O(S)$ be the time and space complexities of the enhanced set-based DA scheme in Algorithm 5, respectively. Then, similar to Algorithm 4, the time complexity of De-SAG in Algorithm 5 is upper bounded by $O(n^2 \log n + T)$ and the space complexity of De-SAG is also $O(S)$. Again, the actual time complexity of De-SAG depends on the particular enhanced structure-based DA attack.

5.4.2 Evaluation

In this subsection, we evaluate the performance of De-SAG and compare it with state-of-the-art structure-based DA attacks.

5.4.2.1 Evaluation Setting

Since De-SAG has two implementations depending on the enhanced structure-based DA attacks, we compare De-SAG with the latest user-based DA scheme proposed in [69], denoted by VLDB14, and the latest set-based DA scheme proposed in [63],

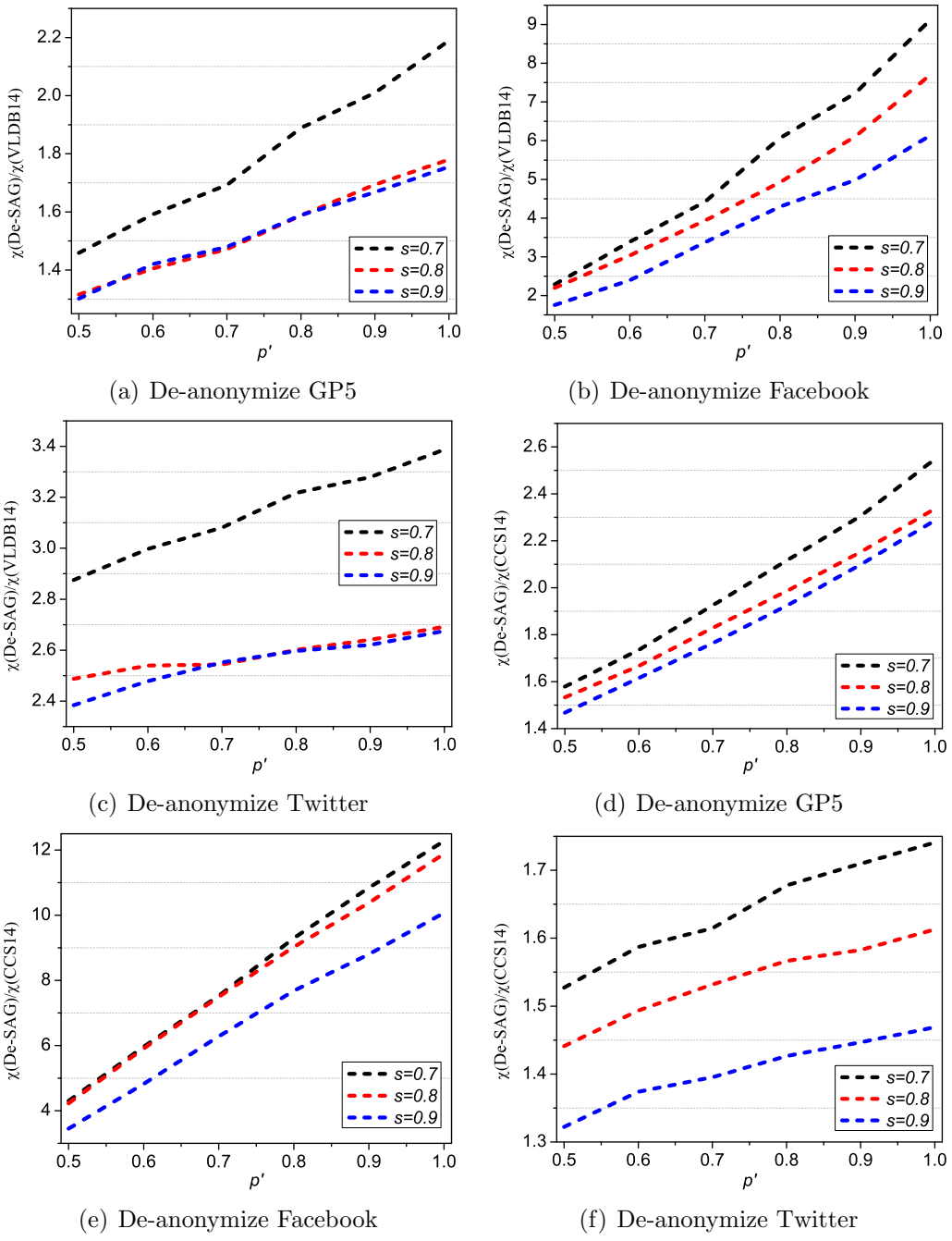
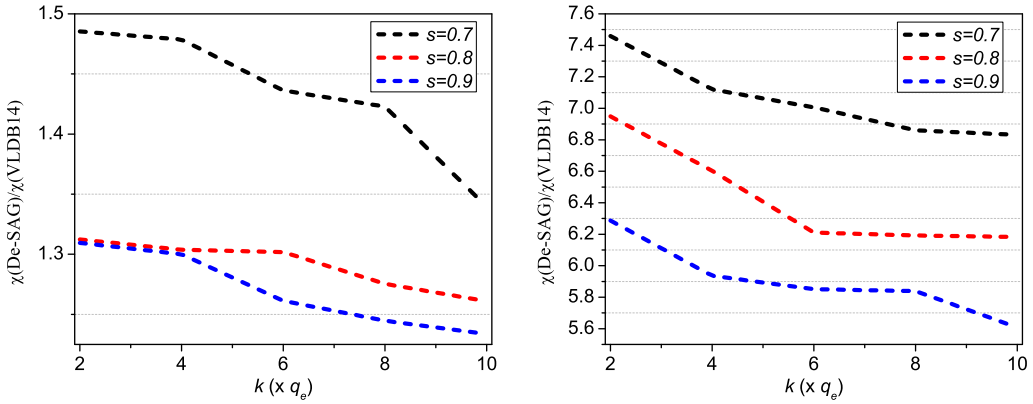


Figure 25: De-SAG Evaluation (vs p'). Default setting: $q' = q'' = 0$ and $c = 0.5$.

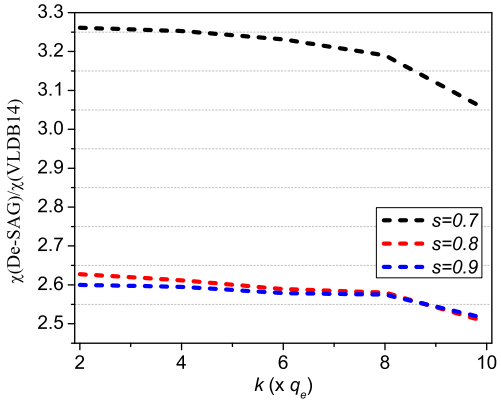
denoted by CCS14.

To conduct the evaluation, we employ three SAG datasets from Table 10: G-P5, Facebook, and Twitter, and follow the following methodology. First, given a

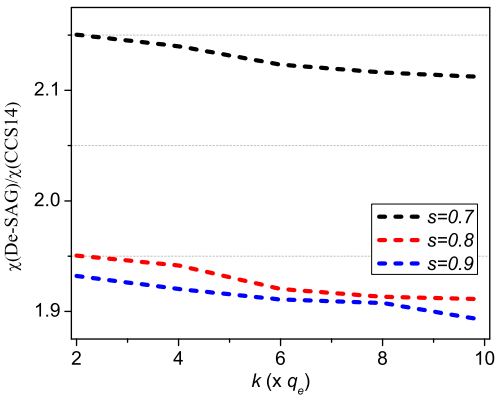


(a) De-anonymize GP5

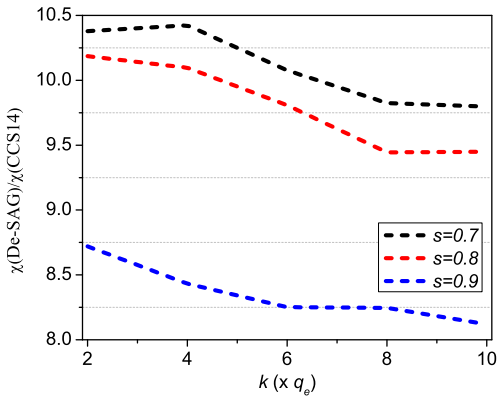
(b) De-anonymize Facebook



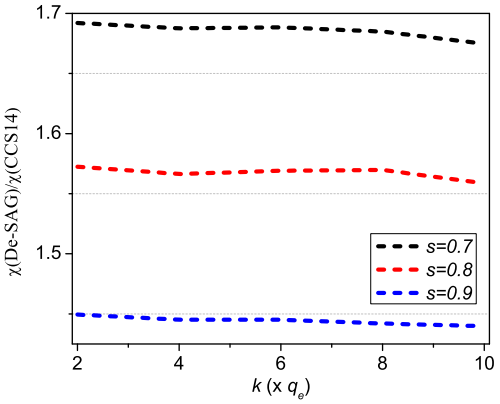
(c) De-anonymize Twitter



(d) De-anonymize GP5



(e) De-anonymize Facebook



(f) De-anonymize Twitter

Figure 26: De-SAG evaluation (vs q'). Default setting: $p' = p'' = 0.8$ and $c = 0.5$.

raw dataset $G = (V, E, A, W)$ (i.e., GP5, Facebook, and Twitter here), we obtain the anonymized graph $G' = (V', E', A', W')$ and the auxiliary graph $G'' = (V'', E'', A'', W'')$ according to the parameter setting of each group of evaluations.

When constructing (V', E') and (V'', E'') from G, we employ the same technique as in [63, 69] for fairness and accuracy. Specifically, we let $V' = V'' = V$ and E' and E'' are random subsets of E with each link in E appearing in E'/E'' with probability s, i.e., $\Pr(l_{ij} \in E' | l_{ij} \in E) = \Pr(l_{ij} \in E'' | l_{ij} \in E) = s$. For A' and A'', we assume $A' = A'' = A$ according to our model. We also determine W' and W'' according to our data model. Specifically, similar to the evaluation setting of our theoretical anonymity analysis, we set $\Pr(a_{ij} \in W' | a_{ij} \in W) = p' = \Pr(a_{ij} \in W'' | a_{ij} \in W) = p''$ and $\Pr(a_{ij} \in W' | a_{ij} \notin W) = q' = \Pr(a_{ij} \in W'' | a_{ij} \notin W) = q''$. Second, we employ VLDB14, CCS14, and De-SAG to de-anonymize G' leveraging G'', respectively. The *successful DA rate* of each DA algorithm is defined as $\chi(\cdot) = \frac{n_c}{n}$, where n_c is the number of users that have been successfully de-anonymized and $n = |V|$ is the total number of users in an anonynized dataset.

As summarized in Section 2, VLDB14 is a seed-based attack and CCS14 is a seed-free attack. Therefore, in our evaluation, we feed VLDB14 50 seed mappings, which are the top-50 users in G with respect to node degree. For other parameters, we specify them in each group of evaluations.

5.4.2.2 Results

In Fig.25, we show the impacts of p' on the performance of VLDB14, CCS14, and De-SAG when de-anonymizing GP5, Facebook, and Twitter. Specifically, we show the change of $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ with respect to the increase of p' in Fig.25 (a)-(c) and the change of $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$ with respect to the increase of p' in Fig.25 (d)-(f), respectively. We analyze Fig.25 as follows.

1. When p' increases, both $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$ increase under different s. This is because when p' increases, more attribute information appears in both the anonymized graph and the auxiliary graph, i.e., the users in G' and G'' have

more common attributes (which implies that the users in G' and G'' have better attribute similarity in Algorithms 4 and 5). Then, χ(De-SAG) increases since more attribute information is available for DA, followed by the increase of $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$. Note that this result is also consistent with our theoretical analysis and experimental evaluation in Section 5.3: the increase of p' implies the decrease of the anonymity of G'.

2. When de-anonymizing GP5, Facebook, and Twitter, on average, the successful DA rate of De-SAG is 1.63, 4.63, and 2.75 times of that of VLDB14 respectively, and is 1.94, 7.79, and 1.53 times of that of CCS14 respectively. This demonstrates that the attribute information is very powerful in enhancing existing structure-based DA attacks, which further confirms our attribute-based anonymity analysis (the attribute information can significantly reduce the anonymity SAG data).

3. In most of the scenarios, De-SAG leads to more improvements compared to VLDB14 and CCS14 for smaller s than larger s. For instance, when de-anonymizing Facebook employing De-SAG and VLDB14 (Fig.25 (b)), on average, $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})} = 5.42$ when $s = 0.7$, $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})} = 4.65$ when $s = 0.8$, and $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})} = 3.82$ when $s = 0.9$; and when de-anonymizing Twitter employing De-SAG and CCS14 (Fig.25 (f)), on average, $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})} = 1.64$ when $s = 0.7$, $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})} = 1.54$ when $s = 0.8$, and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})} = 1.41$ when $s = 0.9$. This is because a small s implies that less links in E appeare in E' and E'', followed by less structural similarity between G' and G''. Therefore, the structure-based DA attacks VLDB14 and CCS14 will have a performance degradation. On the other hand, the attributes associated with users can provide relatively more useful information for successful DA.

Leveraging GP5, Facebook, and Twitter, we show the impacts of q' on the performance of VLDB14, CCS14, and De-SAG in Fig.26, where q' is defined as $q' = k \cdot q_e$ ($k = 2, \cdots, 10$, and $q_e = \frac{(1-p')M}{|nN-M|}$ which is the same as in Section 5.3). Specifically, the impacts of q' on $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ are shown in Fig.26 (a)-(c), and the impacts of q' on $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$ are shown in Fig.26 (d)-(f), respectively. We analyze the results in Fig.26 as follows.

1. When q' increases, both $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$ decrease under different s. For instance, when de-anonymizing GP5 employing De-SAG and VLDB14 in the case of $s = 0.7$ (Fig.26 (a)), $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ is decreased from 1.49 to 1.34 when q' is increased from $2q_e$ to $10q_e$; and when de-anonymizaing GP5 employing De-SAG and CCS14 in the case of $s = 0.7$ (Fig.26 (d)), $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$ is decreased from 2.15 to 2.11 when q' is increased from $2q_e$ to $10q_e$. This is because, as indicated in Section 5.3, with the increase of q', more fake user-attribute links will be added to G' and G'', and thus the benefit of employing the attribute information for DA is decreased, followed by the decrease of χ(De-SAG). Then, both $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})}$ and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})}$ decrease. This is consistent with our analysis and evaluation in Section 5.3.

2. As in Fig.25, when de-anonymizing GP5, Facebook, and Twitter, on average, the successful DA rate of De-SAG is 1.33, 6.46, and 2.77 times of that of VLDB14 respectively, and is 1.99, 9.41, and 1.57 times of that of CCS14 respectively. This demonstrates that De-SAG can significantly improve existing structure-based DA attacks by taking account both the structure and the attribute information.

3. Given q', similar to that in Fig.25, the improvements of De-SAG over VLDB14/CCS14 is higher for smaller s in most of the scenarios. For instance, when de-anonymizing Facebook (Fig.26 (b) and (e)), on average, $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})} = 7.06$ when $s = 0.7$, $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})} = 6.43$ when $s = 0.8$, and $\frac{\chi(\text{De-SAG})}{\chi(\text{VLDB14})} = 5.9$ when

$s = 0.9$; and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})} = 10.1$ when $s = 0.7$, $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})} = 9.8$ when $s = 0.8$, and $\frac{\chi(\text{De-SAG})}{\chi(\text{CCS14})} = 8.33$ when $s = 0.9$. Again, this is due to the fact that the attributes associated with users can relatively provide more information for DA when less structural information is available.

5.4.3 Discussion

Based on our analysis and the evaluation results, De-SAG can significantly improve the performance of existing structure-based DA attacks by taking account both structure and attribute information. Therefore, in graph data sharing/publishing research, it is also important to protect the user-attribute relationships in addition to protecting the graph structure. However, to the best of our knowledge, most, if not all, of the existing graph anonymization techniques only consider to anonymize graph structure [97, 122, 159]. Hence, we plan to conduct SAG data anonymization research in the future by considering both the graph structure and the user-attribute relationships. Our attribute-based anonymity analysis and evaluation are expected to shed light on such research.

5.5 Chapter Summarization

In this chapter, we study the impacts of the attribute information (non-PII) on the privacy of SAG data both theoretically and experimentally. First, we conduct an attribute-based anonymity analysis for SAG data. By careful quantification, we explicitly obtain the correlation between the graph anonymity and the associated attribute information. Through numerical and real world data-based evaluations, we validate our analysis and show that the attribute information may cause significant graph anonymity loss. Subsequently, according to our attribute-based anonymity analysis, we propose a novel DA framework, namely De-SAG, to graph data, which takes account both graph structure and attribute information. By extensive evaluation, we demonstrate that De-SAG can significantly improve the performance of

state-of-the-art DA attacks. Our attribute-based anonymity analysis and DA framework are expected to fill the gap in understanding the actual privacy vulnerability of graph data and further shed light on future graph anonymization and DA research.

CHAPTER VI

AUD: QUANTIFYING THE ANONYMITY-UTILITY-DE-ANONYMITY OF GRAPH DATA

6.1 Introduction

Several graph anonymization, De-Anonymization (DA), and DA quantification techniques have been proposed. However, there are still some *important yet open* problems in this area, such as: *is there correlation between the anonymity, utility, and de-anonymity of graph data? if the correlation exists, how can it be quantified? what is the performance of state-of-the-art anonymization and DA techniques and is there room for improvement?* Understanding these open problems are important for users and researchers. On one hand, it can help users and researchers understand how much utility is preserved after applying an anonymization scheme, what is the achievable data anonymity, what is the achievable de-anonymity given an auxiliary graph, as well as the correlation between anonymity, utility, and de-anonymity. On the other hand, it can also help users and researchers evaluate the performance of state-of-the-art anonymization and DA techniques compared to the achievable theoretical anonymity and de-anonymity.

To address the aforementioned problems, in this chapter, we quantify and evaluate the *Anonymity-Utility-De-anonymity* (AUD) for graph data under both the mathematical Erdős-Rényi (ER) model and a general data model. We also apply our AUD quantification to evaluate the performance of state-of-the-art anonymization and DA techniques. Specifically, our contributions can be summarized as follows.

- We introduce three metrics to measure the anonymity, utility, and de-anonymity

of anonymized graph data, respectively. Based on these metrics, we conduct a comprehensive quantification of the correlation of graph anonymity, utility, and de-anonymity under both the mathematical ER model and a general data model. To the best of our knowledge, *this is the first work on quantifying the AUD correlation of graph data and providing close-forms to explicitly demonstrate such correlation.*

- Based on our correlation quantification, we conduct a large scale evaluation on the AUD of real world graph data leveraging 12 datasets that are generated from various computer systems and services. Our results demonstrate that the achievable anonymity/de-anonymity of graph data depends on multiple factors, e.g., the utility carried by the data, the quality of the employed auxiliary data.

- Based on our AUD quantification, we evaluate the performance of state-of-the-art anonymization and DA techniques. Interestingly, we find that there is still a significant room for state-of-the-art DA techniques to be improved. More importantly, *for the first time*, our results also explicitly and quantitatively indicate the improvement room. For instance, when using the latest seed-free DA attack ODA ([63], Chapter 3) to de-anonymize a Facebook dataset (64K users, 0.82M edges) that is anonymized by the state-of-the-art DP-based anonymization technique [118, 122], our evaluation shows that more than 83.4% theoretically de-anonymizable users cannot be correctly de-anonymized by ODA, i.e., ODA can be significantly improved.

The rest of this chapter is organized as follows. In Section 6.2, we provide the system model. We conduct the AUD correlation quantification under the ER model and a general model in Sections 6.3 and 6.4, respectively. In Section 6.5, we evaluate the AUD of real world graph data. In Section 6.6, we conduct the AUD quantification based evaluation of the performance of existing anonymization and DA techniques.

We conclude this chapter in Section 6.7.

6.2 System Model and Definitions

In this section, we provide the system model and the formal definitions of *utility*, *de-anonymity*, and *anonymity*.

6.2.1 Utility

First, we model the *raw data* (e.g., social network data, email networks, contact graphs) for sharing/publishing by a graph $G^r = (V^r, E^r)$, where $V^r = \{1, 2, \cdots\}$ and $E^r = \{e_{i,j} | i, j \in V^r\}$ characterize the set of users and the set of relationships among users in the dataset respectively. Let $|V^r| = n$, i.e., the number of users is n. When sharing/publishing G^r, it is first anonymized by an arbitrary anonymization technique denoted by Π. Let $G^a = (V^a, E^a) = \Pi(G^r)$ be the anonymized graph. Without loss of generality, we assume $V^a = V^r$ (this is consistent with existing anonymization techniques [38, 53, 88, 97, 118, 122, 131, 142, 152, 158, 160]).

As shown in [38, 53, 88, 118, 122, 131, 142, 152, 158, 160], the utility of G^a can be measured by many perspectives, e.g., *degree distribution*, *joint degree distribution*, *cluster coefficient*, *network resilience*, application-based utilities, etc. These metrics demonstrate the utility of the data from different perspectives. However, a general metric does not exist. On the other hand, we notice that existing utility metrics depend highly on how structurally/topologically similar G^r and G^a are. Therefore, we define an *edge-based general utility metric* μ. The objectives of defining μ are the following: *consistent with existing utility metrics*; *sufficiently general to characterize the correlation between the raw and anonymized graphs*; and *mathematically tractable when quantifying the correlation of anonymity, utility, and de-anonymity.* For $G^a = \Pi(G^r)$, it is defined as

$$\mu(G^a) = \frac{|E^a \cap E^r| + |\overline{E^a} \cap \overline{E^r}|}{|E^U|}, \tag{168}$$

where $|\cdot|$ is the *cardinality* of a set, E^U is the *universal set* of all the possible edges that can be formed among users in V^r, $\overline{E^a} = E^U \setminus E^a = \{e_{i,j} | e_{i,j} \notin E^a\}$, and $\overline{E^r} = E^U \setminus E^r$. To be more accurate, we further define $\mu_1 = \frac{|E^a \cap E^r|}{|E^a|}$ and $\mu_0 = \frac{|\overline{E^a} \cap \overline{E^r}|}{|\overline{E^a}|}$. Then,

$$\mu(G^a) = \frac{\mu_1 |E^a| + \mu_0 |\overline{E^a}|}{|E^U|}. \tag{169}$$

Therefore, when $\mu_1 = \mu_0$, we have $\mu(G^a) = \mu_1 = \mu_0$. From the definition of $\mu(G^a)$, it measures the degree of G^a on preserving the structure (both the existing and the non-existing relationships) of G^r. We further experimentally demonstrate the performance of μ_Π in Section 6.5.

6.2.2 De-anonymity

To de-anonymize G^a, as in [61, 63, 104, 108, 127], we assume an *auxiliary graph* $G^u = (V^u, E^u)$ is available to the adversary. In reality, G^u can be obtained through multiple means, e.g., online crawling, data mining and aggregation, government publishing, third-party applications [63, 104, 108, 127]. Without loss of generality, we assume $V^u = V^r = V^a = V$ (this is a common assumption in existing analysis [61, 63, 69, 113]). When $V^u \neq V^a$, we can (i) either apply our analysis to the overlap users of V^a and V^u, or (ii) simply redefine $V^a = V^a \cup (V^u \setminus V^a)$ and $V^u = V^u \cup (V^a \setminus V^u)$ without changing E^a or E^u. Since G^u and G^r/G^a characterize the relationship of a same group of users, it is reasonable to assume G^u and G^r/G^a are correlated with each other. For instance, let G^r be an email network and G^u be an auxiliary Google+ graph of the same user set V. Then, for two users Alice and Bob in V, if they have a connection in G^r, they are also more likely to have a connection in Google+. To characterize this correlation between G^r and G^u, we statistically define

$$\Pr(e_{i,j} \in E^u | e_{i,j} \in E^r) = \tau, \tag{170}$$

and

$$\Pr(e_{i,j} \in E^u | e_{i,j} \notin E^r) = \gamma, \tag{171}$$

i.e., statistically, an edge appearers in G^u with probability τ when it is also appeared in G^r while with probability γ when it does not appear in G^r.

To be consistent with existing work [61, 63, 69, 113], we mathematically define a DA attack as a mapping

$$\sigma : V^a \to V^u. \tag{172}$$

Specifically, $\sigma := \{(i, \sigma(i))|i \in V^a, \sigma(i) \in V^u\}$. To simplify the discussion, a mapping $(i, \sigma(i))$ is correct when $i = \sigma(i)$ and incorrect otherwise. Given σ, let ω be the number of incorrect mappings in σ. Then, the ratio of successfully de-anonymized users in G^a under σ is defined as

$$\beta_\sigma = \frac{n - \omega}{n}. \tag{173}$$

Let $\mathbb{S}$ be the set of all the possible DA schemes. Since $|V^u| = |V^a| = n$, evidently, we have $n!$ possible mappings from V^a to V^u, i.e., $|\mathbb{S}| = n!$. The *de-anonymity* of G^a is defined as

$$\beta(G^a) = \max\{\beta_\sigma|\sigma \in \mathbb{S}\}. \tag{174}$$

From the definition, the de-anonymity of G^a is measured by the maximum ratio of users that can be successfully de-anonymized. Intuitively, for an anonymized graph G^a, its practical de-anonymity depends on multiple factors, e.g., the correlation between the anonymized graph and the auxiliary graph. Therefore, it is difficult, if not possible, to derive the exact $\beta(G^a)$ for an arbitrary G^a. A practical quantification would seek to understand the de-anonymity of G^a relative to the utility carried by G^a. Toward this objective, we quantify the the lower bound of the de-anonymity of G^a given the utility of G^a and an auxiliary graph in our AUD quantification.

6.2.3 Anonymity

We employ an information theoretical approach to define the anonymity of G^r/G^a given Π and σ, which is similar to that in [108]. For a user $i \in V^a$ and $\forall j \in V^u$, let

$p_{i,j}$ be the probability of the event that i *is mapped to (de-anonymized as)* j *in a DA scheme* σ *(i.e.,* $(i,j) \in \sigma$*) and this mapping is a correct DA, i.e.,* $j = \sigma(i)$ *(*i *and* j *correspond to the same user).* For instance, if σ randomly and uniformly maps each $i \in V^a$ to any user $j \in V^u$, then we have $p_{i,j} = \frac{1}{n}$, i.e., i is successfully de-anonymized with probability $\frac{1}{n}$ under σ.

Based on the definition of $p_{i,j}$, we define $\mathbf{P}_{\Pi,\sigma}(i) = \{p_{i,j} | j \in V^u\}$ to be the mapping probability distribution of i under σ. Then, using information theory, the uncertainty of de-anonymizing i can be measured by entropy

$$H(i) = -\sum_{j \in V^u} p_{i,j} \log p_{i,j}. \tag{175}$$

Evidently, when $\mathbf{P}_{\Pi,\sigma}(i) = \{p_{i,j} = \frac{1}{n} | j \in V^u\}$ (i is mapped to any user in V^u randomly and uniformly), i.e., the successful DA probability of i is $p_{i,j} = \frac{1}{n}$ for $\forall j \in V^u$ under Π, $H(i)$ reaches its maximum value $\log n$. In this scenario, an anonymization scheme Π is optimal from the perspective of protecting the privacy of i. On the other hand, if $\mathbf{P}_{\Pi,\sigma}(i) = \{p_{i,1} = 0, \cdots, p_{i,i-1} = 0, p_{i,i} = 1, p_{i,i+1} = 0, \cdots, p_{i,n} = 0\}$, i.e., the probability that i is successfully de-anonymized is 1, $H(i)$ reaches its minimum value 0. In this scenario, Π cannot protect the anonymity/privacy of i at all, i.e., the DA scheme σ can successfully break the privacy of i.

Based on $H(i)$, we can quantify the uncertainty of de-anonymizing G^a, denoted by $H(G^a)$, by the average entropy of all the users [108], i.e.,

$$H(G^a) = \frac{1}{n} \sum_{i \in V^a} H(i). \tag{176}$$

Let $H_{\max}(G^a)$ be the maximum entropy that G^a can be achieved. Since $\max\{H(i)\} = \log n$, we have

$$H_{\max}(G^a) = \log n. \tag{177}$$

Here, if $H(G^a) = H_{\max}(G^a) = \log n$, G^a achieves the optimal anonymity. Then, the

anonymity of G^a is defined as

$$\alpha(G^a) = \frac{H(G^a)}{H_{\max}(G^a)} = \frac{H(G^a)}{\log n}, \tag{178}$$

which measures how optimal G^a is on achieving uncertainty. Specifically, $\alpha(G^a) \in [0, 1]$, where 1 implies G^a achieves the best anonymity while 0 implies no anonymity at all.

From the anonymity definition, it is measured by the uncertainty of the process of de-anonymizing G^a. When studying the AUD correlation, our objective is to quantify the upper bound of the achievable $\alpha(G^a)$ relative to the utility preserved in G^a and the available auxiliary graph G^u.

In the remainder of this chapter, we use $\mu = \mu(G^a)$, $\beta = \beta(G^a)$, and $\alpha = \alpha(G^a)$ for convenience of discussion. In addition, for the lowercase parameter x, we define $\overline{x} = 1 - x$, e.g., when $x = \mu$, $\overline{x} = \overline{\mu} = 1 - \mu$.

6.3 AUD Quantification: ER Model

In this section, we quantify the AUD of graph data under the *Erdős-Rényi* (ER) model. We extend our quantification to the general scenario in the next section.

6.3.1 Preliminaries

Suppose G^r follows the ER model $G(n, p)$, i.e., there are n users in G^r and $\forall i, j \in V^r$, the edge $e_{i,j}$ appeares in E^r with probability p ($\Pr(e_{i,j} \in E^r) = p$). When sharing/publishing G^r, it is first anonymized by an arbitrary anonymization scheme Π and the obtained anonymized graph is G^a. $\forall i \in V^a$, its neighborhood is defined as $N_i^a = \{j | \exists e_{i,j} \in E^a\}$. Similarly, $\forall i \in V^u$, we define $N_i^u = \{j | \exists e_{i,j} \in E^u\}$.

Given $\sigma : V^a \to V^u$, to measure the quality of the mapping $(i, j) \in \sigma$, similarly as in [61, 63, 113], we define a *Neighborhood Difference Function* (NDF)

$$\Delta_{\sigma:(i,j)} = |(\bigcup_{v \in N_a^i} \{\sigma(v)\}) \setminus N_j^u| + |(\bigcup_{v \in N_u^j} \{\sigma^{-1}(v)\}) \setminus N_i^a|, \tag{179}$$

i.e., $\Delta_{\sigma:(i,j)}$ counts the neighborhood difference of $i \in V^a$ and $j \in V^u$ under σ. Then, to measure σ, we define the NDF of σ as

$$\Delta_\sigma = \sum_{i \in V^a} \Delta_{\sigma:(i,\sigma(i))}. \tag{180}$$

6.3.2 Quantification

We quantify the AUD of an anonymized graph in this subsection. First, we quantify the NDF of a given σ. Given Π, G^a, G^u, and σ, let μ be the utility of G^a, $q_c(\mu) = p\mu_1\overline{\tau}+\overline{p}\cdot\overline{\mu_0}\cdot\overline{\gamma}+p\overline{\mu_1}\tau+\overline{p}\mu_0\gamma$, and $q_{i,c}(\mu) = (p\mu_1+\overline{p}\cdot\overline{\mu_0})(p\overline{\tau}+\overline{p}\cdot\overline{\gamma})+(p\overline{\mu_1}+\overline{p}\mu_0)(p\tau+\overline{p}\gamma)$. Then, we show the NDF of σ in Lemma 2.

Lemma 2. *If there are ω incorrect mappings in σ,*

$$\Delta_\sigma \underset{n \to \infty}{\sim} B(\binom{n-\omega}{2}, q_c(\mu)) + B(\omega(n-\omega) + \binom{\omega}{2}, q_{i,c}(\mu)), \tag{181}$$

where $B(\cdot,\cdot)$ denotes a binomial variable.

Proof: To facilitate our analysis, we denote the correctly and incorrectly de-anonymized users by V_c and V_ω, respectively. Furthermore, let E_c, E_ω, and $E_{c,\omega}$ be all the possible edges among the users in V_c, among the users in V_ω, and between the users in V_c and V_ω, respectively. Hence, $|V_c| = n-\omega$ and $|V_\omega| = \omega$.

To quantify Δ_σ, we employ an *edge-based quantification* technique. Specifically, we consider the following three cases.

Case 1: $e_{i,j} \in E_c$. In this case, $e_{i,j}$ causes a Neighborhood Difference (ND) when it appears in one graph (G^a/G^u) while not in the other one (G^u/G^a). Specifically, the probability of causing this ND is

$$p\mu_1\overline{\tau} + \overline{p}\cdot\overline{\mu_0}\cdot\overline{\gamma} + p\overline{\mu_1}\tau + \overline{p}\mu_0\gamma = q_c(\mu). \tag{182}$$

Therefore, the NDs caused by the edges in E_c is a *binomial variable* $B(\binom{n-\omega}{2}, q_c(\mu))$.

Case 2: $e_{i,j} \in E_{c,\omega}$. In this case, $e_{i,j}$ is an edge connecting a correctly de-anonymized user (suppose it is i) and an incorrectly de-anonymized user (suppose

it is j). Hence, $e_{i,j}$ will cause a ND if only one of $e_{i,j}$ and $e_{i,\sigma(i)}$. Let $q_{i,c}$ denote the probability that $e_{i,j}$ causes a ND. Then,

$$(p\mu_1 + \overline{p} \cdot \overline{\mu_0})(p\overline{\tau} + \overline{p} \cdot \overline{\gamma}) + (p\overline{\mu_1} + \overline{p}\mu_0)(p\tau + \overline{p}\gamma) = q_{i,c}(\mu). \tag{183}$$

Therefore, the NDs caused by the edges in $E_{c,\omega}$ is also a binomial variable $B(\omega(n - \omega), q_{i,c}(\mu))$.

Case 3: $e_{i,j} \in E_\omega$. In this case, $e_{i,j}$ connects two incorrectly de-anonymized users i and j. We further partition the edges in E_ω into two subcases. First, under σ, if $\sigma(i) = j$ and $\sigma(j) = i$, then $e_{i,j}$ is a *transposition edge*. In this subcase, $e_{i,j}$ causes a ND with probability $q_c(\mu)$. Furthermore, the number of transposition edges in E_ω is at most $\omega/2$. In this second subcase, $e_{i,j}$ is called a *non-transposition edge* and it causes a ND with probability $q_{i,c}(\mu)$. Therefore, the NDs caused by the edges in E_ω is the sum of two binomial variables $B(t, q_c(\mu)) + B(\binom{\omega}{2} - t, q_{i,c}(\mu))$, where t is the number of transposition edges in $E_{c,\omega}$ and $t \leq \omega/2$.

In summary, we have

$$\Delta_\sigma \sim B(\binom{n-\omega}{2}, q_c(\mu)) + B(\omega(n-\omega), q_{i,c}(\mu)) + B(t, q_c(\mu)) + B(\binom{\omega}{2} - t, q_{i,c}(\mu)) \tag{184}$$

$$= B(\binom{n-\omega}{2} + t, q_c(\mu)) + B(\omega(n-\omega) + \binom{\omega}{2} - t, q_{i,c}(\mu)) \tag{185}$$

$$\underset{n \to \infty}{\simeq} B(\binom{n-\omega}{2}, q_c(\mu)) + B(\omega(n-\omega) + \binom{\omega}{2}, q_{i,c}(\mu)). \tag{186}$$

□

Based on Lemma 2, we can quantify the correlation of the *utility*, *anonymity*, and *de-anonymity* of an anonymized graph as shown in Theorem 22.

Theorem 22. *Let* $f(\mu) = \frac{(q_{i,c}(\mu) - q_c(\mu))^2}{8(q_{i,c}(\mu) + q_c(\mu))}$ *be a* utility function *depending on the utility of* G^a *and* ω *be the number of possibly incorrectly de-anonymized users in a DA scheme. Then, when* $q_{i,c}(\mu) > q_c(\mu)$ *and* $f(\mu) = \Omega(\frac{2\ln n + 1}{\omega n - \omega^2/2 - \omega/2})$, *(i)*

$$\beta = \Omega(\frac{n-\omega}{n}); \tag{187}$$

and (ii)

$$\alpha = O(\frac{\omega}{n}\log_n \omega). \tag{188}$$

Proof: (i) First, we prove the first conclusion. Let $\mathbb{S}'$ be the set of all the possible DA schemes from V^a to V^u (including the scheme that correctly de-anonymizes all the users in V^a). Evidently, $|\mathbb{S}'| = n!$, which is *countable* and *enumerable*. Therefore, to prove our conclusion, it is sufficient to prove that *given μ, we can find a DA scheme $\sigma \in \mathbb{S}'$ such that at most ω users are incorrectly de-anonymized under σ.*

Let σ be *the scheme in $\mathbb{S}'$ such that σ induces the least NDF*, i.e., $\Delta_\sigma = \min\{\Delta_\Im | \Im \in \mathbb{S}'\}$. Intuitively, σ can be found in $\mathbb{S}'$ in finite time by a *brute-force* searching algorithm, and thus σ is a deterministic DA scheme. In addition, let $\sigma^* \in \mathbb{S}$ be the *optimal* DA scheme such that all the users are correctly de-anonymized (note that, we do not actually know which scheme is σ^* in $\mathbb{S}'$ till now although it exists). Furthermore, let $\sigma' \in \mathbb{S}$ be another scheme such that ω ($\omega \in [2, n]$) users are incorrectly de-anonymized. Based on Lemma 2, we have

$$\Delta_{\sigma^*} \underset{n \to \infty}{\sim} B(\binom{n}{2}, q_c(\mu)), \tag{189}$$

and

$$\Delta_{\sigma'} \underset{n \to \infty}{\sim} B(\binom{n-\omega}{2}, q_c(\mu)) \tag{190}$$

$$+B((\omega)(n-\omega) + \binom{\omega}{2}, q_{i,c}(\mu)). \tag{191}$$

Let

$$X \sim B(\binom{\omega}{2} + \omega(n-\omega), q_c(\mu)) \tag{192}$$

and

$$Y \sim B(\binom{\omega}{2} + \omega(n-\omega), q_{i,c}(\mu)). \tag{193}$$

Further, let λ_X and λ_Y be the expectation values of X and Y, respectively. Since $q_{i,c}(\mu) > q_c(\mu)$, $\lambda_Y > \lambda_X$. Hence, according to the Pedarsani-Grossglauser Lemma [113],

$$\Pr(\Delta_{\sigma^*} \geq \Delta_{\sigma'}) \tag{194}$$

$$= \Pr(\lambda_X \geq \lambda_Y) \tag{195}$$

$$\leq 2\exp(-\frac{(\lambda_Y - \lambda_X)^2}{8(\lambda_Y + \lambda_X)}) \tag{196}$$

$$= 2\exp(-(\binom{\omega}{2} + \omega(n-\omega)) \cdot \frac{(q_{i,c}(\mu) - q_c(\mu))^2}{8(q_{i,c}(\mu) + q_c(\mu))}) \tag{197}$$

$$= 2\exp(-(\binom{\omega}{2} + \omega(n-\omega)) \cdot f(\mu)) \tag{198}$$

$$\leq 2\exp(-2\ln n - 1) \tag{199}$$

$$\leq \frac{1}{n^2}. \tag{200}$$

Then, according to the Borel-Cantelli Lemma, $\Pr(\Delta_{\sigma^*} \geq \Delta_{\sigma'}) \underset{n \to \infty}{\to} 0$, i.e., $\Pr(\Delta_{\sigma^*} < \Delta_{\sigma'}) \underset{n \to \infty}{\to} 1$. Furthermore, considering that $\Delta_\sigma = \min\{\Delta_\Im | \Im \in \mathbb{S}'\}$ and $\sigma^* \in \mathbb{S}'$, we have $\Delta_\sigma \leq \Delta_{\sigma^*} < \Delta_{\sigma'}$. Therefore, we conclude that the DA scheme σ has at most ω incorrect mappings. It follows that $\beta \geq \beta_\sigma = \Omega(\frac{n-\omega}{n})$.

(ii) Now, we prove the second conclusion. From the proof of the first conclusion, given G^u and σ, at least $n - \omega$ users can be successfully de-anonymized. Let V_c^a be the set of users that are successfully de-anonymized under σ and $V_c^u = \{\sigma(i) \in V^u | i \in V_c^a\}$. Then, we determine the mapping probability distribution of each user in V^a. $\forall i \in V^a$, if $i \in V_c^a$, $\mathbf{P}_{\Pi,\sigma}(i) = \{0, \cdots, 0, p_{i,i} = 1, 0, \cdots, 0\}$; if $i \in V_c^a$, $\mathbf{P}_{\Pi,\sigma}(i) = \{p_{i,j} = 0 | j \in V_c^u\} \cup \{p_{i,j} \in [0,1] | j \in V^u \setminus V_c^u\}$. Therefore, $H(i) = 0$ if $i \in V_c^a$ and $H(i) \leq \log\omega$ if $i \in V^a \setminus V_c^a$. It follows that $H(G^a) \leq \frac{1}{n} \cdot \omega\log\omega$ and thus $\alpha = \frac{H(G^a)}{\log n} = O(\frac{\omega}{n}\log_n \omega)$. □

Remarks. In Theorem 22, we quantify the correlation between μ, β, and α. From the quantification results, the lower bound of the utility function $f(\mu)$ is defined by a decreasing function of ω (the number of possible incorrect mappings). When

parameter ω increases, a looser condition is required for the utility function $f(\mu)$, followed by a lower de-anonymity β and a higher anonymity α of G^a are achievable. On the other hand, if higher de-anonymity is expected (i.e., lower anonymity can be achieved by G^a), a stricter condition is required on $f(\mu)$. Furthermore, from the proof of Theorem 22, When the specified conditions on $q_c(\mu)$, $q_{i,c}(\mu)$, and $f(\mu)$ are satisfied, a DA scheme σ that correctly de-anonymizes at least $n-\omega$ users can be found by a brute-force searching algorithm. Although the searching algorithm has a time complexity of $O(n!)$, which makes it computationally infeasible in reality, practical heuristics/approximation-optimization based DA attacks can be designed, e.g., [63, 104, 127]. Therefore, the significance of our quantification is to help users/researchers understand the theoretical correlation of anonymized graph data's utility, anonymity, and de-anonymity; and thus improve the anonymization/DA research.

6.4 AUD Quantification: In General

In the previous section, we quantified the correlation of data utility μ, de-anonymity β, and anonymity α of G^a, given Π, G^u, and σ. However, as indicated in [61, 63], it is seldom to see, if ever seen, that real world graph data follow the ER model. The reason is that a graph under the ER model has a Poisson degree distribution, while real world graph data may follow any degree distribution, e.g., the power-law distribution. Although the ER model is more likely to be a theoretical model, the quantification under the ER model can still shed light on the quantification of utility, de-anonymity, and anonymity under a practical model. In this section, we quantify the correlation of μ, β, and α under a general model, where the graph can have an arbitrary degree distribution.

We assume $G^r(V^r, E^r)$ can follow an arbitrary degree distribution. Let $m_r = |E^r|$. The *graph density* of G^r is defined as $\rho = \frac{m_r}{|E^U|} = \frac{2m_r}{n(n-1)}$. Let $\phi_c(\mu) = \mu_1\overline{\tau} + \overline{\mu_1}\tau$, $\phi_{i,c}(\mu) = \mu_1(\rho\overline{\tau} + \overline{\rho}\cdot\overline{\gamma}) + \overline{\mu_1}(\rho\tau + \overline{\rho}\gamma)$, $\psi_c(\mu) = \overline{\mu_0}\cdot\overline{\gamma} + \mu_0\gamma$, and $\psi_{i,c}(\mu) = \overline{\mu_0}(\rho\overline{\tau} +$

$\overline{\rho} \cdot \overline{\gamma}) + \mu_0(\rho\tau + \overline{\rho}\gamma)$. Before quantifying the correlation of μ, β, and α, we first use Lemma 3 to quantify the NDF of a DA scheme σ, which has ω incorrect mappings.

Lemma 3. *Let* $\theta_{\min} = \min\{\phi_c(\mu), \psi_c(\mu)\}$, $\theta_{\max} = \max\{\phi_c(\mu), \psi_c(\mu)\}$, $\tau_{\min} = \min\{\phi_{i,c}(\mu), \psi_{i,c}(\mu)\}$, *and* $\tau_{\max} = \max\{\phi_{i,c}(\mu), \psi_{i,c}(\mu)\}$. *If there are* ω *incorrect mappings in* σ,

$$\Delta_\sigma \geq B(\binom{n-\omega}{2}, \theta_{\min}) + B(\omega(n-\omega) + \binom{\omega}{2}, \tau_{\min}) \tag{201}$$

and

$$\Delta_\sigma \leq B(\binom{n-\omega}{2}, \theta_{\max}) + B(\omega(n-\omega) + \binom{\omega}{2}, \tau_{\max}). \tag{202}$$

Proof: To facilitate our analysis, we basically employ the same notations as in Lemma 2. Specifically, we denote the correctly and incorrectly de-anonymized users by V_c and V_ω, respectively. Furthermore, let E_c, E_ω, and $E_{c,\omega}$ be all the possible edges among the users in V_c, among the users in V_ω, and the users in V_c and V_ω, respectively. Then, we conduct our proof by quantifying the ND caused by each edge in E^U. Let $\Delta_{e_{i,j}}$ be the ND caused by $e_{i,j}$. Specifically, we have the following six cases.

Case 1: $e_{i,j} \in E^r \cap E_c$. In this case, $e_{i,j}$ appears in G^r and i and j are correctly de-anonymized under σ. Therefore, $e_{i,j}$ causes a ND if it is appeared in one graph (G^a/G^u) while not in the other one (G^u/G^a). It follows that $\Delta_{e_{i,j}} \sim B(1, \mu_1\overline{\tau}+\overline{\mu_1}\tau) = B(1, \phi_c(\mu))$.

Case 2: $e_{i,j} \in E^r \cap E_{c,\omega}$. In this case, $e_{i,j}$ appeares in G^r and exactly one of i and j is incorrectly de-anonymized. Suppose j is incorrectly de-anonymized. Then, a ND is caused if $e_{i,j}$ appeares in G^a while $e_{i,\sigma(j)}$ does not appeare in G^u and vice versa. Therefore, $\Delta_{e_{i,j}} \sim B(1, \mu_1(\rho\overline{\tau} + \overline{\rho} \cdot \overline{\gamma}) + \overline{\mu_1}(\rho\tau + \overline{\rho}\gamma)) = B(1, \phi_{i,c}(\mu))$.

Case 3: $e_{i,j} \in E^r \cap E_\omega$. In this case, $e_{i,j}$ appeares in G^r and both i and j are incorrectly de-anonymized under σ. To quantify the ND caused by $e_{i,j}$, we consider two subcases. In the first subcase, if $e_{i,j}$ is a transposition edge under σ, then $\Delta_{e_{i,j}} \sim B(1, \phi_c(\mu))$; otherwise, $\Delta_{e_{i,j}} \sim B(1, \phi_{i,c}(\mu))$.

Case 4: $e_{i,j} \in \overline{E^r} \cap E_c$. In this case, $e_{i,j}$ is not an edge in G^r and both i and j are correctly de-anonymized under σ. Again, $e_{i,j}$ will case a ND if it appears in G^a/G^u while not in G^u/G^a. Therefore, $\Delta_{e_{i,j}} \sim B(1, \overline{\mu_0} \cdot \overline{\gamma} + \mu_0\gamma) = B(1, \psi_c(\mu))$.

Case 5: $e_{i,j} \in \overline{E^r} \cap E_{c,\omega}$. In this case, $e_{i,j}$ does not appear in E^r and exactly one of i and j is incorrectly de-anonymized under σ. Therefore, $\Delta_{e_{i,j}} \sim B(1, \overline{\mu_0}(\rho\overline{\tau} + \overline{\rho} \cdot \overline{\gamma}) + \mu_0(\rho\tau + \overline{\rho}\gamma)) = B(1, \psi_{i,c}(\mu))$.

Case 6: $e_{i,j} \in \overline{E^r} \cap E_\omega$. In this case, $e_{i,j}$ does not appear in G^r and both i and j are incorrectly de-anonymized. Then, if $e_{i,j}$ is a transposition edge under σ, $\Delta_{e_{i,j}} \sim B(1, \psi_c(\mu))$; otherwise, $\Delta_{e_{i,j}} \sim B(1, \psi_{i,c}(\mu))$.

Let t_1 and t_2 be the numbers of transposition edges in $E^r \cap E_\omega$ and $\overline{E^r} \cap E_\omega$, respectively. Then, $t_1 + t_2 \leq \omega/2$. In summary,

$$\Delta_\sigma \sim B(|E^r \cap E_c| + t_1, \phi_c(\mu)) + B(|E^r \cap E_{c,\omega}| + |E^r \cap E_\omega| - t_1, \phi_{i,c}(\mu)) \tag{203}$$

$$+ B(|\overline{E^r} \cap E_c| + t_2, \psi_c(\mu)) + B(|\overline{E^r} \cap E_{c,\omega}| + |\overline{E^r} \cap E_\omega| - t_2, \psi_{i,c}(\mu)) \tag{204}$$

$$\underset{n \to \infty}{=} B(|E^r \cap E_c|, \phi_c(\mu)) + B(|E^r \cap E_{c,\omega}| + |E^r \cap E_\omega|, \phi_{i,c}(\mu)) \tag{205}$$

$$+ B(|\overline{E^r} \cap E_c|, \psi_c(\mu)) + B(|\overline{E^r} \cap E_{c,\omega}| + |\overline{E^r} \cap E_\omega|, \psi_{i,c}(\mu)) \tag{206}$$

Therefore,

$$B(|E_c|, \theta_{\min}) + B(|E_{c,\omega}| + |E_\omega|, \tau_{\min}) \leq \Delta_\sigma \leq B(|E_c|, \theta_{\max}) + B(|E_{c,\omega}| + |E_\omega|, \tau_{\max}), \tag{207}$$

i.e.,

$$\Delta_\sigma \geq B(\binom{n-\omega}{2}, \theta_{\min}) + B(\omega(n-\omega) + \binom{\omega}{2}, \tau_{\min}) \tag{208}$$

and

$$\Delta_\sigma \leq B(\binom{n-\omega}{2}, \theta_{\max}) + B(\omega(n-\omega) + \binom{\omega}{2}, \tau_{\max}). \tag{209}$$

□

In Lemma 3, we derive the lower and upper bounds of Δ_σ for a given σ. Based on Lemma 3, we quantify the correlation of μ, β, and α under a general data model in Theorem 23.

Theorem 23. *Let* $g(\mu) = \frac{(\min\{\phi_{i,c}(\mu),\psi_{i,c}(\mu)\}-\max\{\phi_c(\mu),\psi_c(\mu)\})^2}{8(\min\{\phi_{i,c}(\mu),\psi_{i,c}(\mu)\}+\max\{\phi_c(\mu),\psi_c(\mu)\})} = \frac{(\tau_{\min}-\theta_{\max})^2}{8(\tau_{\min}+\theta_{\max})}$ *be a* utility function *depending* μ*, and* ω *be the number of possibly incorrectly de-anonymized users in a DA scheme. Then, when* $\tau_{\min} > \theta_{\max}$ *and* $g(\mu) = \Omega(\frac{2\ln n+1}{\omega n-\omega^2/2-\omega/2})$*, (i)*

$$\beta = \Omega(\frac{n-\omega}{n}); \tag{210}$$

and (ii)

$$\alpha = O(\frac{\omega}{n}\log_n \omega). \tag{211}$$

Proof Sketch: Basically, this theorem can be proven by applying similar techniques as in Theorem 22. For convenience, we use the same notations as in Theorem 22. Specifically, let $\mathbb{S}'$ be the set of all the possible DA schemes, $\sigma \in \mathbb{S}'$ such that $\Delta_\sigma = \min\{\Delta_\Im | \Im \in \mathbb{S}'\}$, $\sigma^* \in \mathbb{S}'$ be the optimum DA scheme, and $\sigma' \in \mathbb{S}'$ be a DA scheme with ω incorrect mappings. Evidently, σ is a DA scheme that can be found by a brute force algorithm in finite time $O(n!)$.

(i) To prove the first conclusion, it is sufficient to prove that *under* σ*, at least* $n-\omega$ *users are successfully de-anonymized.* Based on Lemma 3, we have

$$B(\binom{n}{2}, \theta_{\min}) \leq \Delta_{\sigma^*} \leq B(\binom{n}{2}, \theta_{\max}). \tag{212}$$

and

$$\Delta_{\sigma'} \geq B(\binom{n-\omega}{2}, \theta_{\min}) + B(\omega(n-\omega) + \binom{\omega}{2}, \tau_{\min}). \tag{213}$$

Let $X \sim B(\omega(n-\omega)+\binom{\omega}{2}, \theta_{\max})$ and $Y \sim B(\omega(n-\omega)+\binom{\omega}{2}, \tau_{\min})$. Furthermore, let λ_X and λ_Y be the expectation values of X and Y, respectively. Since $\theta_{\max} < \tau_{\min}$, we

Table 11: Data statistics.

Name	Type	n	m	ρ	d_{avg}
Wiki (WK)	WikiTalk Data	2.4M	5M	1.63E-6	3.9
Gnutella (GT)	P2P Network Data	36.7K	88.3K	1.32E-4	4.8
YouTube (YT)	Social Networks	1.1M	3M	4.64E-6	5.3
Oregon (OG)	Autonomous Systems	11.5K	32.7K	4.98E-4	5.7
Brightkite (BK)	Location-based Social Networks	58K	.2M	1.32E-4	7.5
Gowalla (GW)	Location-based Social Networks	.2M	1M	4.92E-5	9.7
Enron (EN)	Email Data	36.7K	.2M	3.19E-4	10.7
Skitter (SK)	Autonomous Systems	1.7M	11.1M	7.73E-6	13.1
Facebook (FB)	Social Networks	64K	.82M	4.02E-4	25.64
Google+ (G+)	Social Networks	4.7M	90.8M	8.24E-6	38.7
Twitter (TW)	Social Networks	.5M	14.9M	1.20E-4	54.8
Flickr (FL)	Social Networks	80.5K	5.9M	1.82E-3	146.56

have

$$\Pr(\Delta_{\sigma^*} \geq \Delta_{\sigma'}) \leq \Pr(X \geq Y) \tag{214}$$

$$= 2\exp(-(\binom{\omega}{2} + \omega(n-\omega)) \cdot g(\mu)) \tag{215}$$

$$\leq 2\exp(-2\ln n - 1) \tag{216}$$

$$\leq \frac{1}{n^2}. \tag{217}$$

Therefore, $\Pr(\Delta_{\sigma^*} < \Delta_{\sigma'}) \underset{n\to\infty}{\to} 1$. Then, we have $\Delta_\sigma \leq \Delta_{\sigma^*} < \Delta_{\sigma'}$, which implies that at most ω users are incorrectly de-anonymized under σ and thus $\beta \geq \beta_\sigma \geq \frac{n-\omega}{n}$, i.e., $\beta = \Omega(\frac{n-\omega}{n})$.

(*ii*) Based on the above proof, this conclusion can be proven using the similar technique as in Theorem 22. □

Remarks. From Theorem 23, the correlation of μ, β, and α under a general model is similar to that under the ER model. However, they are different with respect to *required conditions* and *generality/applicability*. Fundamentally, to achieve the same anonymity/de-anonymity, the conditions under the general model (specified by $g(\mu)$,

$\tau_{\min}$, and $\theta_{\max}$) are stricter than that under the ER model (specified by $f(\mu)$, $q_c(\mu)$, and $q_{i,c}(\mu)$). On the other hand, the quantification in Theorem 22 is dedicated for graphs under the ER model while the quantification in Theorem 23 is applicable to graphs following any distribution.

6.5 Utility Metric and AUD Evaluation

In this section, we evaluate our AUD quantification using real world graph datasets.

6.5.1 Datasets

In the evaluation, we employ 12 real world graph datasets, which are generated from various computer systems: Social Networks, Location-based Social Networks, Email networks, WikiTalk networks, P2P networks, and Autonomous Systems. All the datasets are now publicly available and can be found at Berkeley Datasets [15], Stanford SNAP [16], and ASU Datasets [1]. We show the basic statistical information of the 12 datasets in Table 11, where n, m, ρ, and d_{avg} denote the number of users, the number of edges, the graph density, and the average degree of each user, respectively.

We briefly introduce each dataset as follows.

- *Wiki* (WK). The Wiki dataset is a graph consisting of Wikipedia users and their Wikipedia Talk (WikiTalk) relationship.
- *Gnutella* (GT). The Gnutella dataset is a Peer-to-Peer (P2P) file sharing graph, where nodes represent hosts/users and edges represent sharing connections among users.
- *YouTube* (YT). YouTube is a video-sharing service. The YouTube dataset is a social graph of YouTube users and their relationships.
- *Oregon* (OG). The Oregon dataset is a route-view Internet topology graph representing the connections (edges) of routers (nodes) in an Autonomous System (AS).

- *Brightkite* (BK). The Brightkite dataset is a location-based social network graph, where nodes represent users and edges represent the friendships among users.

- *Gowalla* (GW). The Gowalla dataset is also a location-based social network graph representing the friendships among users.

- *Enron* (EN). The Enron dataset is an email communication graph, where nodes represent users and edges represent the email communication relationships among users.

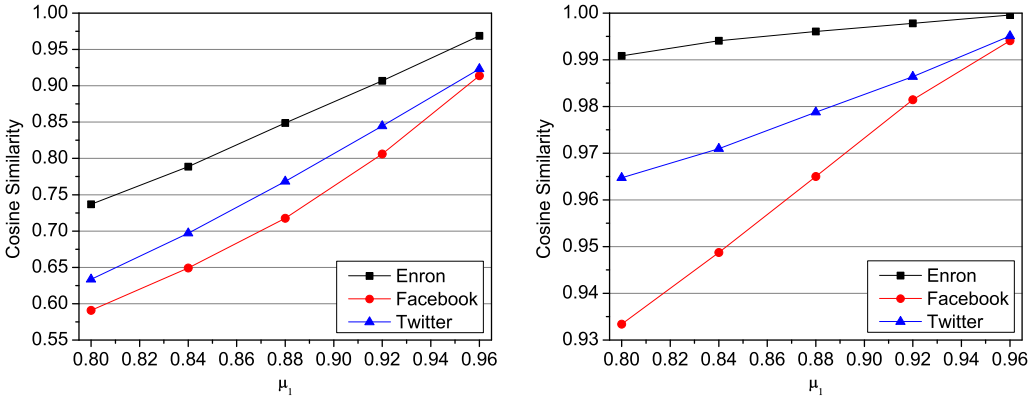
- *Skitter* (SK). Similar as *Oregon*, the Skitter dataset is also an Internet topology graph.

- *Facebook* (FB). Facebook is one of the most popular online social networking service. The Facebook dataset is a social network graph where nodes represent users and edges represent friendships.

- *Google+* (G+). Google+ is a social network and a social layer for Google services. The Google+ dataset is a social network graph of Google+ users and their connections.

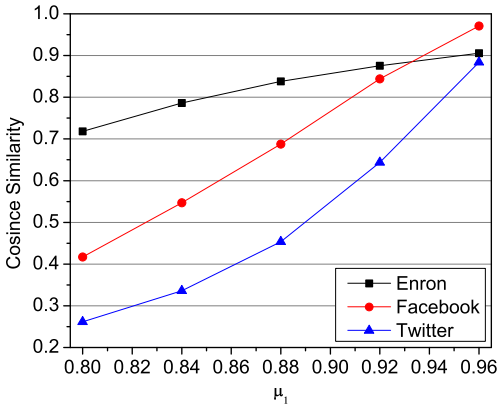
- *Twitter* (TW). Twitter is an online social networking service that enables users to send and read short 140-character messages called "tweets". The Twitter dataset is a social network graph of Twitter users and their following relationships.

- *Flickr* (FL). Flickr is an image hosting website. The Flickr dataset is a social network graph of Flickr users and their friendships.



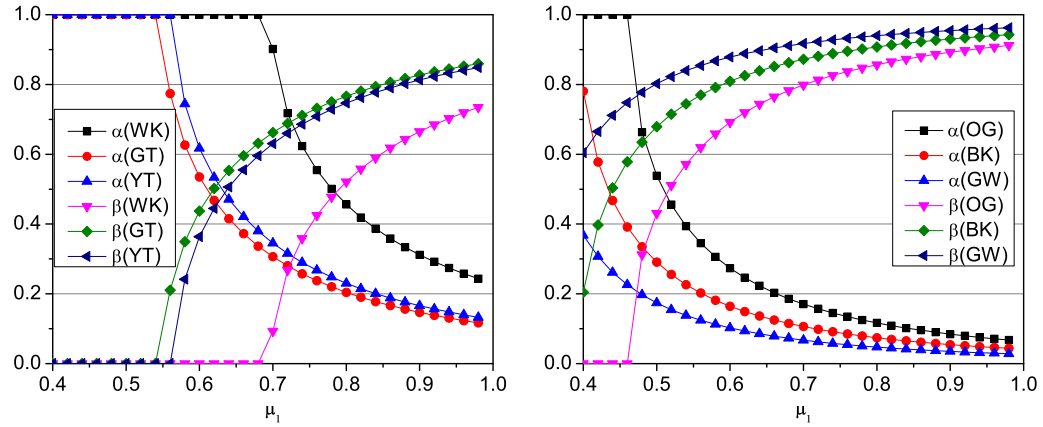
(a) Degree similarity

(b) PL similarity



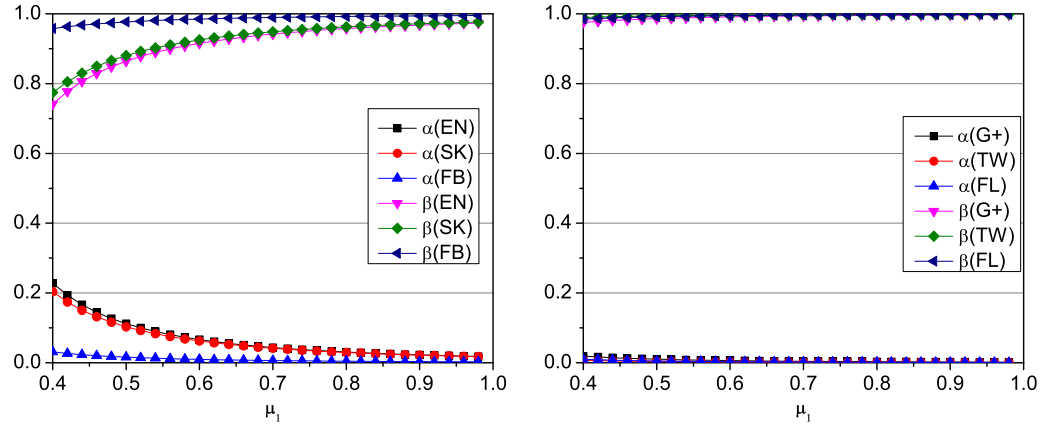
(c) CC similarity

Figure 27: The performance of the utility metric μ.



(a) Wiki (WK), Gnutella (GT), and YouTube (YT)

(b) Oregon (OG), Brightkite (BK), and Gowalla (GW)



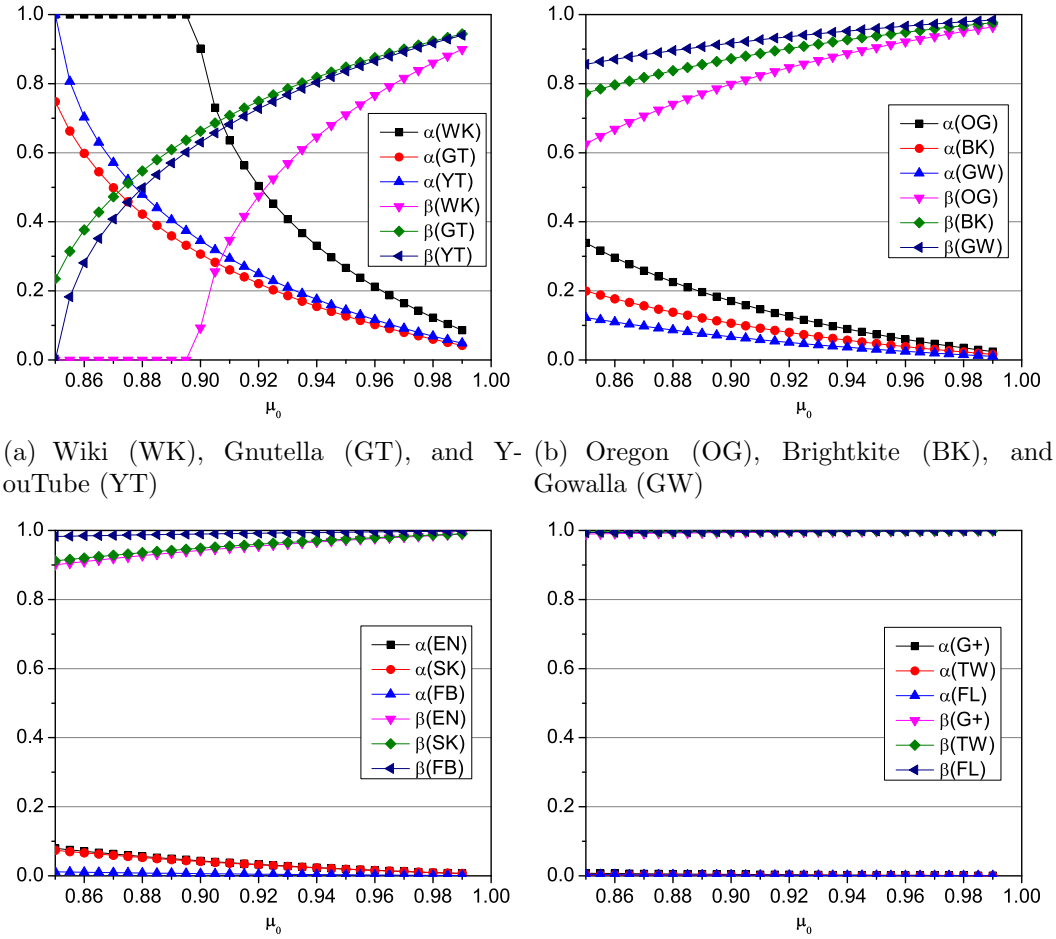
(c) Enron (EN), Skitter (SK), and Facebook (FB)

(d) Google+ (G+), Twitter (TW), and Flickr (FL)

Figure 28: AUD vs. μ_1.

6.5.2 Performance of the Utility Metric μ

Before evaluating our AUD quantification, we first examine the performance of our utility metric μ. According to the definition of μ, it measures the performance of G^a on preserving the structure (both the existing and the non-existing relationships) of G^r. Furthermore, since μ is defined based on μ_1 and μ_0, we examine the effectiveness of μ by evaluating the utility of G^a with respect to different μ_1 and μ_0. Due to the space limitation, here, we employ three datasets Enron, Facebook, and Twitter as example datasets for the evaluation, and the evaluated utilities of G^a are *Degree distribution*



(a) Wiki (WK), Gnutella (GT), and YouTube (YT)

(b) Oregon (OG), Brightkite (BK), and Gowalla (GW)

(c) Enron (EN), Skitter (SK), and Facebook (FB)

(d) Google+ (G+), Twitter (TW), and Flickr (FL)

Figure 29: AUD vs. μ_0.

(Deg), *Path Length distribution* (PL), and *Cluster Coefficient distribution* (CC)[1]. The reason we choose to evaluate Deg, PL, and CC is because they are the most fundamental graph utilities and most of the other graph utilities (e.g., infectiousness, reliable email, secure routing, community property, and influence propagation) are highly dependent on them [137, 152].

Evaluation Methodology. The evaluation methodology is as follows: (i) given

[1]In SecGraph, 12 graph utilities (e.g., degree distribution, joint degree distribution, path length, closeness centrality distribution) and 7 application utilities (e.g., influence maximization, community detection, secure routing) are implemented [20, 62]. Using SecGraph, it is straightforward to evaluate the effectiveness of μ with respect to these graph and application utilities. Basically, our utility metric is positively correlated with the graph and application utilities in SecGraph.

μ_1 and μ_0 and a raw dataset G^r, we employ the *Rand Add/Del* anonymization algorithm in [152] to anonymize G^r (by deleting existing edges and adding new edges) such that the obtained anonymized graph G^a has utility of μ_1 and μ_0; (*ii*) compute the Deg, PL, and CC utilities of both G^r and G^a; and (*iii*) compute the *cosine similarity* of each utility of G^r and G^a.

Results. We show the evaluation results in Fig.27, where when changing μ_1 in each evaluation, we set $\mu_0 = 1 - \frac{\overline{\mu_1} \cdot |E^a|}{|E^U \backslash E^a|}$ (note that, μ_0 is an *increasing* function of μ_1)[2]. From Fig.27, we have the following two observations.

First, with the increase of our utility metric μ_1/μ_0, all the three fundamental graph utilities Deg, PL, and CC are also increasing, which demonstrates that *our utility metric is consistent with existing utility metrics*. The reason is because μ_1 and μ_0 measures the degree of G^a to preserve the existing and non-existing relationships of G^u. When G^a and G^u share more common relationships, they are more structurally similar followed by high utility of G^a. Furthermore, based on Theorems 22 and 23, μ_1/μ_0 also enables our AUD quantification tractable. Therefore, our utility metric is effective.

Second, the changing magnitude of PL is smaller than the other two utilities with the increase of μ_1/μ_0. This is because the graph diameters of Enron, Facebook, and Twitter are 10, 10, and 7 respectively, which are relatively small. Therefore, the impact of anonymization (adding/deleting edges) to PL is also relatively small. On the other hand, when μ_1/μ_0 is small, a significant number of relationships in G^r have been changed in G^a. Since Deg and CC are local graph properties, they are more sensitive to local edge changes, i.e., μ_1/μ_0.

[2]The purpose of this setting is to make G^a have relatively similar performance on preserving the existing and non-existing relationships of G^r.

6.5.3 AUD Evaluation

In this subsection, we evaluate the AUD of real world graph datasets (shown in Table 11) based on our AUD quantification.

Evaluation Methodology. For each dataset, it is the raw graph G^r in our evaluation. Then, given μ_1, μ_0, τ, and γ, the structures of both G^a and G^u can be derived from G^r. Finally, we apply our quantification technique in Section 6.4 to quantify the anonymity and de-anonymity of G^a based on G^u. Specifically, according to our proofs in Theorem 22 and Theorem 23, statistically, the optimum DA scheme (mapping) incudes the least NDF. Therefore, after specifying G^a and G^u, we can drive the anonymity and de-anonymity of G^a based on the utility preserved in G^a and G^u (relative to the raw data G^r) using Theorem 23 (the proof of Theorem 23). Note that, here, we are not trying to quantify the exact anonymity/de-anonymity of G^a (and thus, we do not need to seek the optimum DA scheme). Our objective is to derive the upper bound of the achievable anonymity and the lower bound of the achievable de-anonymity with statistical guarantee. Furthermore, in all the evaluations in this subsection, the default parameter settings are $\mu_1 = 0.7$, $\mu_0 = 0.9$, $\tau = 0.75$, and $\gamma = 0.02$.

6.5.3.1 AUD vs. μ

First, we evaluate the anonymity and de-anonymity of the datasets in Table 11 with respect to the utility (characterized by μ_1 and μ_0) preserved by G^a. The results are shown in Fig.28 (changing μ_1) and Fig.29 (changing μ_0), $\alpha(\cdot)$ and $\beta(\cdot)$ are the anonymity and de-anonymity of the corresponding dataset, respectively. From Fig.28 and Fig.29, we have there observations.

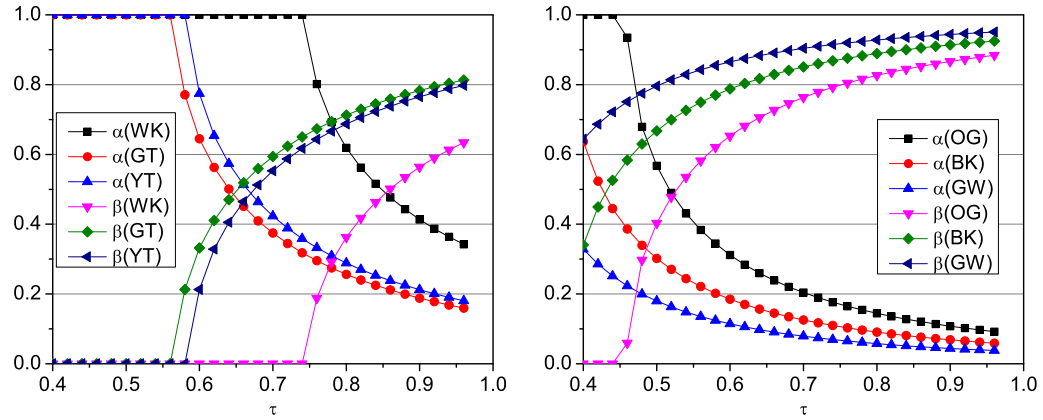
First, when μ_1 (resp., μ_0) increases, the de-anonymity of each dataset increases while the anonymity of each dataset decreases, e.g., in Fig.28 (a), when μ_1 is increased from 0.6 to 0.7, β(YouTube) is increased from 0.364 to 0.63 while α(YouTube) is

decreased from 0.617 to 0.345. This is because μ_1 (resp., μ_0) indicates the degree of G^a on preserving the existing relationships (resp., non-existing relationships) of G^r. A high μ_1 (resp., μ_0) implies that G^a is more structurally similar to G^r, and thus is more structurally similar to G^u (for a given τ and γ). Therefore, more users in G^a are de-anonymizable leveraging the structural similarity between G^a and G^u.

Second, generally, the datasets with high d_{avg} (average degree) are more de-anonymizable (less anonymous) than that with low d_{avg}, e.g., for a given μ_1/μ_0, Facebook ($d_{avg} = 25.64$) is more de-anonymizable than YouTube ($d_{avg} = 5.3$). This is because a higher d_{avg} implies richer local structural information is available in both G^a and G^u for de-anonymizing each user on average. Thus, a user is more likely to be correctly de-anonymized by structure-based DA attacks.

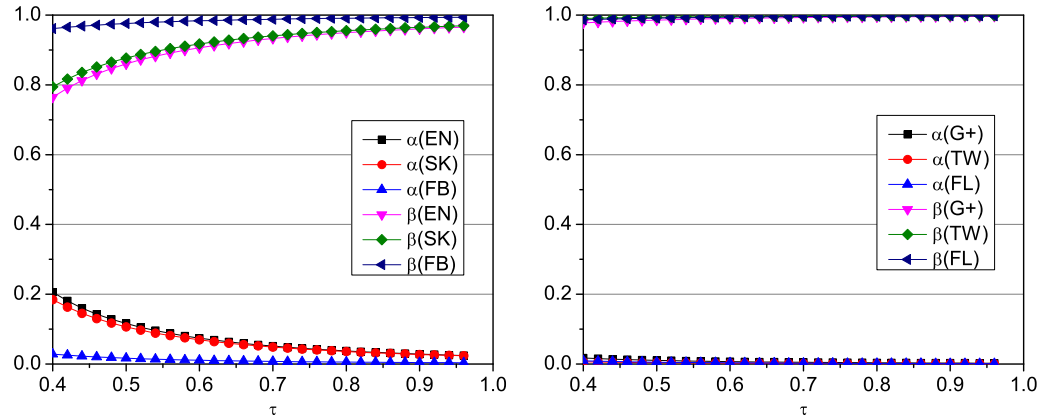
Finally, both the anonymity and the de-anonymity of graph data may exhibit the *percolation phenomena*[3], i.e., when μ_1/μ_0 is below some threshold value, a graph achieves almost perfect anonymity; while when μ_1/μ_0 is above some threshold value, an obvious loss of the anonymity happens. For instance, when μ_1 is increased from 0.46 to 0.48, the anonymity of Oregon is decreased from 0.999 to 0.663. This implies that the actual anonymization/DA performance is sensitive to the utility carried by G^a and the structural similarity between G^a and G^u. Some increase on the similarity of G^a and G^u can induce a significant loss (resp., improvement) of the graph anonymity (resp., de-anonymity).

[3]It has been observed in [61, 104], the number of de-anonymizable users in *seed-based* two-phase DA attacks may exhibit the percolation phenomena with respect to the number of available seeds, i.e., when the number of seeds is below some threshold value, only a few users can be correctly de-anonymized; while when the number of seeds is above some threshold value, a significant portion of users are de-anonymizable.



(a) Wiki (WK), Gnutella (GT), and YouTube (YT)

(b) Oregon (OG), Brightkite (BK), and Gowalla (GW)



(c) Enron (EN), Skitter (SK), and Facebook (FB)

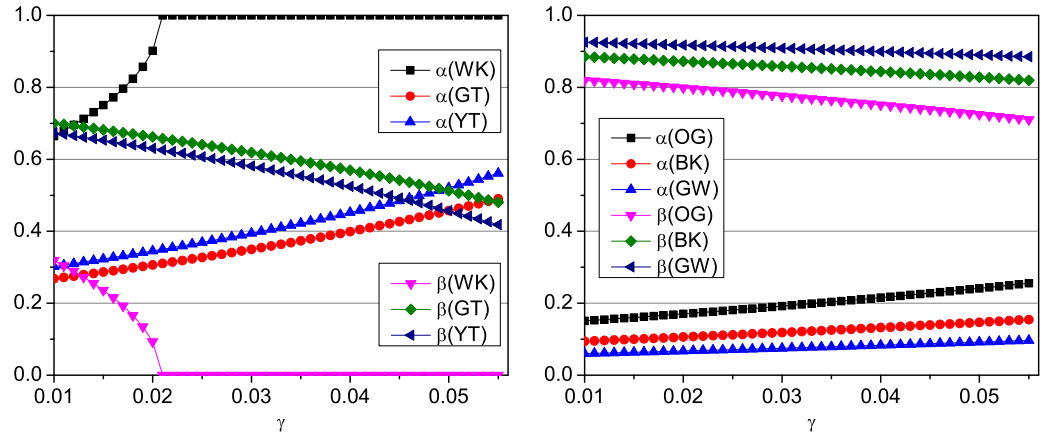
(d) Google+ (G+), Twitter (TW), and Flickr (FL)

Figure 30: AUD vs. τ.

6.5.3.2 AUD vs. τ and γ

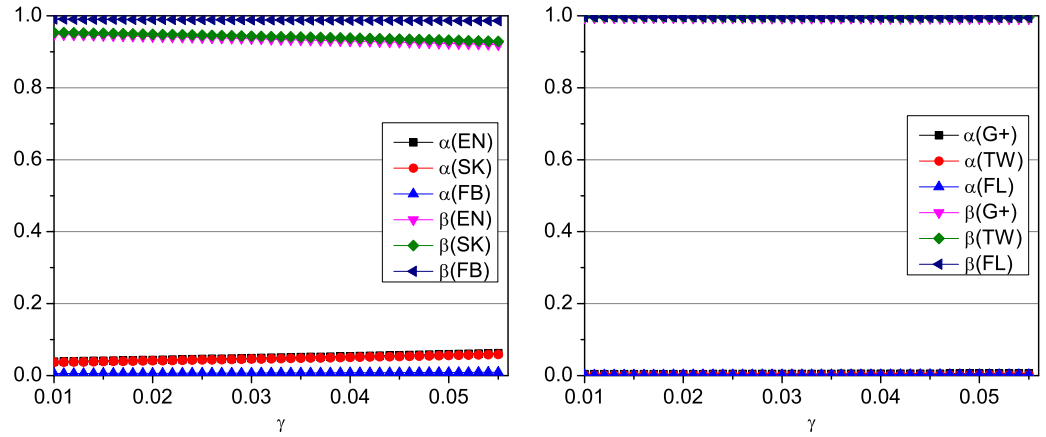
Now, we quantify the AUD of the datasets in Table 11 given different auxiliary graphs, which are characterized by τ and γ. First, when τ increases, the anonymity and de-anonymity of each dataset are shown in Fig.30, from which we have two observations.

First, with the increase of τ, the anonymity (resp., de-anonymity) of each dataset decreases (resp., increases), e.g., when τ is increased from 0.5 to 0.7, α(Oregon) is decreased from 0.57 to 0.204 while β(Oregon) is increased from 0.403 to 0.763. This is because τ indicates how similar G^u and G^r are with respect to the existing relationships in G^r. Thus, a high τ implies G^u is more structurally similar to G^r and



(a) Wiki (WK), Gnutella (GT), and YouTube (YT)

(b) Oregon (OG), Brightkite (BK), and Gowalla (GW)



(c) Enron (EN), Skitter (SK), and Facebook (FB)

(d) Google+ (G+), Twitter (TW), and Flickr (FL)

Figure 31: AUD vs. γ.

thus to G^a (when μ_1, μ_0, and γ are given), followed by G^a is more de-anonymizable by structure-based DA attacks leveraging G^u.

Second, again, the datasets with high d_{avg} are more de-anonymizable than those with low d_{avg}, e.g., Enron ($d_{avg} = 10.7$) is more de-anonymizable than Wiki ($d_{avg} = 3.9$) given a τ. The reason is also the same as we analyzed in Fig.28 and Fig.29. A higher d_{avg} implies richer local structural information is available, which further enables more effective structure-based DA. In addition, similar to that in Fig.28 and Fig.29, the anonymity and de-anonymity of a dataset may exhibit the percolation phenomena with respect to τ, e.g., Wiki, Gnutella, YouTube, and Oregon.

When γ increases, the AUD of the datasets in Table 11 is shown in Fig.31. From Fig.31, we have three observations.

First, when γ increases, the anonymity of each dataset increases while the de-anonymity of each dataset decreases, e.g., when γ is increased from 0.01 to 0.02, α(YouTube) is increased from 0.302 to 0.345 while β(YouTube) is decreased from 0.674 to 0.63. This is because γ indicates the difference of G^r and G^u on users' non-existing relationships. Therefore, a large γ implies more structural difference between G^r and G^u with respect to the non-existing relationships, followed by more structural difference between G^a and G^u. Hence, less structural information can be leveraged to conduct successful DA and the anonymity of G^a is increased.

Second, generally, the anonymity and de-anonymity of datasets with low d_{avg} are more sensitive to the change of γ than that of datasets with high d_{avg}. For instance, when γ is increased from 0.02 to 0.04, α(Gowalla) ($d_{avg} = 9.7$) is increased by 23.6% while α(YouTube) ($d_{avg} = 5.3$) is increased by 31%; at the same time, β(Gowalla) is decreased by 17% and β(YouTube) is decreased by 1.9%. This is because for graphs with lower d_{avg}, the available structural information for de-anonyming each user is relatively less, and thus the structural/edge difference between G^a and G^r has more impacts on the achievable anonymity and de-anonymity.

Finally, similar as the results in Fig.28, Fig.29, and Fig.30, the anonymity and de-anonymity of a dataset may exhibit the percolation phenomena (e.g., Wiki).

6.6 AUD-based Evaluation of State-of-the-Art Anonymization and De-anonymization Techniques

6.6.1 Methodology

In this section, we conduct an AUD-based evaluation of the performance of state-of-the-art graph anonymization and DA techniques. The evaluation methodology is

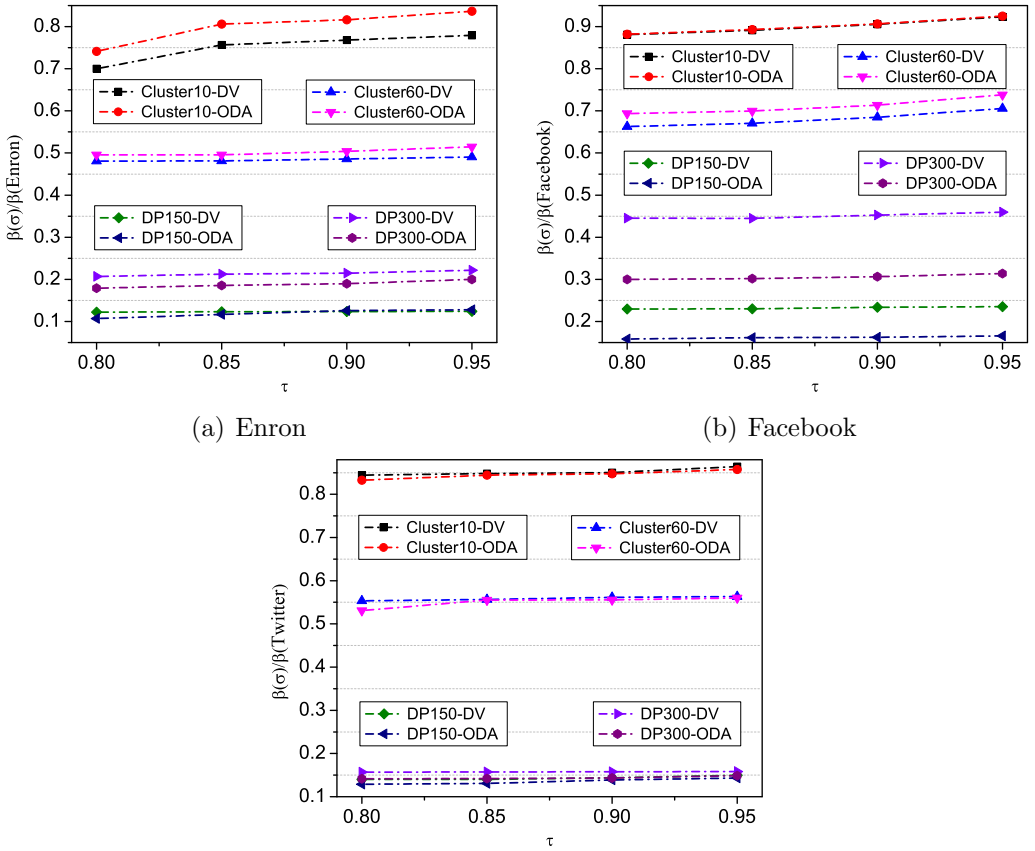
as follows: (i) given some graph datasets, anonymizing these datasets using state-of-the-art anonymization techniques; (ii) employing state-of-the-art DA attacks to de-anonymize the anonymized data and studying the data's practical de-anonymity; (iii) employing our AUD quantification technique to quantify the theoretical de-anonymity (anonymity) of the anonymized data; and (iv) finally, analyzing the practical and theoretical de-anonymity of the anonymized data.

6.6.2 Evaluation Setting

Here, we use three example datasets Enron, Facebook and Twitter as shown in Table 11 for this group of evaluation. The employed anonymization techniques are the latest *cluster-based anonymi-zation technique* [131], denoted by *Cluster*, and the latest *differential privacy-based anonymization technique* [118, 122], denoted by *DP*. As we summarized in Section 2, Cluster is a technique to make the users within a cluster have same local structures, and DP is a technique to make the dK-series of the anonymized graph meet a DP requirement. The employed DA attacks are the *Distance-Vector* (DV) based scheme proposed in [127] and the *Optimization-based DA* (ODA) scheme proposed in [63] (Chapter 3). As summarized in Section 2, DV is a powerful *seed-based* DA attack while ODA is the latest *seed-free* DA attack.

When anonymizing the datasets, the key anonymization parameter for Cluster is the *cluster size* ζ [131] and for DP is the *differential privacy parameter* ξ [118, 122]. Basically, a larger ζ indicates a higher anonymization level for Cluster while a smaller ξ indicates a higher anonymization level for DP. In our evaluation, we consider the scenarios of $\zeta = 10$ and $\zeta = 60$ for Cluster and $\xi = 150$ and $\xi = 300$ for DP (which are similar to the settings in [122, 131]), respectively. For DV, since it requires seeds to bootstrap the DA, we randomly select 50 seed mappings from G^a to G^u in each evaluation. During the DA evaluation, the auxiliary datasets are obtained using a *random edge adding/deleting* process according to the specified τ and γ. In all the

evaluations, we set $\gamma = \frac{\overline{\tau} \cdot |E^r|}{|\overline{E^r}|}$. Furthermore, the required parameters μ_1 and μ_0 for AUD quantification can be obtained according to their definitions given G^r and G^a. For each group of evaluation, it will be repeated 50 times and the result is the average of the 50 runs.



(a) Enron (b) Facebook

(c) Twitter

Figure 32: AUD-based Evaluation of state-of-the-art anonymization and DA techniques.

6.6.3 Resutls

Let σ be a DA attack (e.g., DV and ODA) and n_c be the number of users that are successfully de-anonymized under σ. Then, the *practical de-anonymity* of an anonymized graph under σ is defined as $\beta(\sigma) = \frac{n_c}{n}$. Furthermore, to be consistent with previous notations, we use $\beta(\cdot)$ (e.g., $\beta(\text{Enron})$) to denote the AUD-based de-anonymity of a dataset, i.e., the *theoretical de-anonymity*. Then, we show $\frac{\beta(\sigma)}{\beta(\text{Enron})}$,

$\frac{\beta(\sigma)}{\beta(\text{Facebook})}$, and $\frac{\beta(\sigma)}{\beta(\text{Twitter})}$ under different anonymization/DA scenarios in Fig.32, where "Cluster" and "DP" represent the anonymization algorithms, 10, 60, 150, and 300 represent the anonymization parameters, and DV and ODA represent the DA attacks, respectively. For instance, Cluster10-DV means that the anonymization algorithm applied is Cluster, the anonymization parameter used is 10, and the employed DA attack is DV. From Fig.32, we have there observations.

First, when τ increases, $\frac{\beta(\sigma)}{\beta(\cdot)}$ also has some increase, which implies both DV and ODA can de-anonymize more users regardless of whether the dataset is anonymized by Cluster or DP. The reason is straightforward: a large τ implies G^a and G^u are more structurally similar and thus G^u is more structurally similar to G^a. Therefore, more anonymized users can be successfully de-anonymized based on the structural information (leveraging the structural similarity between G^a and G^u).

Second, for the scenarios of using Cluster as the anonymization algorithm, Cluster60 achieves better anonymity than Cluster10. This is because more users are made structurally similar under Cluster60 than that of Cluster10. However, intuitively, Cluster60 also sacrifices more data utility. Similarly, DP150 achieves better anonymity than DP300 at the cost of sacrificing more data utility. Overall, the datasets anonymized by DP achieves a better anonymity than that of Cluster. This is because DP changes more structural information of G^a than that of Cluster, i.e., the datasets anonymized by Cluster achieves a better utility than DP.

Finally and interestingly, *there is still significant room for state-of-the-art DA techniques to be improved*. From Fig.32, we have $\frac{\beta(\sigma)}{\beta(\cdot)} < 0.95$ in all the scenarios. Specifically, in the scenarios where DP is used, we have $\frac{\beta(\sigma)}{\beta(\text{Enron})} < 0.25$, $\frac{\beta(\sigma)}{\beta(\text{Facebook})} < 0.5$, and $\frac{\beta(\sigma)}{\beta(\text{Twitter})} < 0.16$ for both DV and ODA. Note that, according to our quantification, $\beta(\cdot)$ is only the lower bound of the de-anonymity of an anonymized graph. Therefore, *the practical de-anonymity achieved by state-of-the-art DA attacks are much lower*

than the achievable theoretical de-anonymity. For instance, when using ODA to de-anonymize Facebook anonymized by DP150, $\frac{\beta(\text{ODA})}{\beta(\text{Facebook})} < 0.166$, which implies more than 83.4% theoretically de-anonymizable users cannot be correctly de-anonymized by ODA. Thus, theoretically, significant room exists to improve existing DA attacks in some scenarios.

6.7 Chapter Summarization

In this chapter, we study the correlation of graph data's anonymity, utility, and de-anonymity. Specifically, we conduct the first AUD correlation quantification for anonymized graph data under both the mathematical ER model and a general data model. Based on our quantification, we further conduct a large scale evaluation on the anonymity, utility, and de-anonymity of real world graph data leveraging 12 datasets that are generated from various computer systems and services. Third, we evaluate the performance of state-of-the-art anonymization and DA techniques in terms of our AUD quantification. We find that there is still significant space to improve existing DA attacks, and for the first time, our evaluation results explicitly and quantitatively indicate such possible improvement space. Finally, we discuss to extend and enhance state-of-the-art graph data anonymization and DA evaluation system SecGraph. By adding one quantification module, the functions of SecGraph can be enhanced from multiple perspectives.

CHAPTER VII

SECGRAPH: SECURE GRAPH DATA PUBLISHING/SHARING

7.1 Introduction

As summarized in Chapter 2, to protect graph data's privacy, several anonymization techniques have been proposed to anonymize graph data, which can be classified into six categorizes: Naive ID Removal, Edge Editing (EE) based techniques [152], k-anonymity based techniques [38, 88, 156, 158, 160], Aggregation/Class/Cluster based techniques [30, 53, 131], Differential Privacy (DP) based techniques [117, 118, 122, 137, 142], and Random Walk (RW) based techniques [97]. Fundamentally, these techniques try to protect users' privacy by perturbing the original graph's structure while preserving as much data utility as possible.

Furthermore, following Narayanan and Shmatikov's work [104], many new Structure-based De-Anonymization (SDA, we use DA and SDA interchangeably in this dissertation) attacks on graph data have been proposed, which can be categorized into two classes: *seed-based attacks*, e.g., Narayanan-Shmatikov's attack [104], and *seed-free attacks*, e.g., Ji et al.'s attack [63]. For both types of attacks, the goal is to de-anonymize anonymized users using their uniquely distinguishable structural characteristics.

Surprisingly, although we already have many sophisticated anonymization techniques (e.g., [53, 97, 122, 152, 158]) and powerful SDA attacks (e.g., [63, 69, 102, 104, 108, 112, 127]), whether state-of-the-art anonymization techniques can defend against modern SDA attacks is still an open problem. This is because of the incomplete evaluation of existing anonymization and DA techniques. For anonymization works, they

usually only evaluate the data utility performance of their proposed techniques (although some works provide a theoretical security guarantee, these guarantees usually do not hold due to improper assumptions or incomplete considerations as analyzed in Section 7.4). For DA works, they usually evaluate their attacks' performance without applying state-of-the-art anonymization techniques (e.g., k-anonymity based schemes, DP based schemes) to their test data.

To address the above open problem, we systematically study, implement, and evaluate existing graph data anonymization techniques and DA attacks. Specifically, our main contributions are as follows.

(a) We design and implement a Secure Graph data publishing/sharing (SecGraph) system (available at [20]). SecGraph enables data owners to anonymize their data using state-of-the-art anonymization techniques, measure the anonymized data's graph and application utilities, and comprehensively evaluate their data's actual vulnerability against modern DA attacks. To the best of our knowledge, SecGraph is the first such system publicly available to both academia and industry. More importantly, SecGraph provides the first *uniform platform* that enables researchers to conduct accurate comparative studies of anonymization/DA techniques, and to comprehensively understand the resistance/vulnerability of existing or newly developed anonymization techniques, the effectiveness of existing or newly developed DA attacks, and graph and application utilities of anonymized data.

(b) In SecGraph, we systematically analyze, implement, and evaluate 11 state-of-the-art graph data anonymization schemes and 19 graph and application utility metrics. We also analyze the 11 anonymization schemes with respect to the 19 utility metrics, both analytically and experimentally. The evaluation results demonstrate that most existing anonymization algorithms can partially or conditionally preserve most graph utilities. However, all the anonymization schemes lose one or more application utility.

(c) We summarize and analyze the fundamental properties of existing SDA attacks. Then, we systematically implement and evaluate 15 modern SDA attacks on real-world graph datasets. Our results show that modern SDA attacks are powerful and robust to seed mapping errors. Furthermore, no attack is optimum in all scenarios. The DA performance of an attack depends on the similarity between the anonymized and auxiliary data, graph density, DA heuristics, etc.

(d) We analytically and experimentally evaluate the performance of existing graph data anonymization schemes on defending against modern SDA attacks. We find that existing anonymization techniques are vulnerable to modern SDA attacks. Their degree of vulnerability depends on how much data utility is preserved in the anonymized data.

Abbreviations. For convenient reference, we summarize the used abbreviations in Table 12.

The rest of this Chapter is organized as follows. In Section 7.2, we study existing graph data anonymization schemes and their utility performance. In Section 7.3, we study modern SDA attacks. In Section 7.4, the effectiveness of existing anonymization schemes against modern DA attacks is analyzed. We systematically implement and evaluate SecGraph in Section 7.5 and conclude this chapter in Section 7.6.

7.2 Anonymization and Utility

Generally, an anonymization scheme can be evaluated from two perspectives: *data utility preservation* and *resistance to DA attacks*. However, most, if not all, existing graph anonymization works have not been significantly evaluated with respect to their utility or resistance to DA attacks. On one hand, most existing graph anonymization works only conducted limited evaluations on their utility preservation, e.g., degree distribution, path length distribution, which are insufficient to understand their value

Table 12: Abbreviations and acronyms.

Terms	SDA	Structure-based De-anonymization
	DA	De-anonymization
	SF	Seed-Free
Anonymizatoin	EE	Edge Editing
	DP	Differential Privacy
	RW	Random Walk
	k-NA	k-Neighborhood Anonymity
	k-DA	k-Degree Anonymity
	k-auto	k-automorphism
	k-iso	k-isomorphism
Utility metrics	Deg.	Degree
	JD	Joint Degree
	ED	Effective Diameter
	PL	Path Length
	LCC	Local Clustering Coefficient
	GCC	Global Clustering Coefficient
	CC	Closeness Centrality
	BC	Betweenness Centrality
	EV	Eigenvector
	NC	Network Constraint
	NR	Network Resilience
	Infe.	Infectiousness
	RX	Role extraction
	RE	Reliable Email
	IM	Influence Maximization
	MINS	Minimum-sized Influential Node Set
	CD	Community Detection
	SR	Secure Routing
	SD	Sybil Detection
De-anonymization	DV	Distance Vector [127]
	RST	Randomized Spanning Tress [127]
	RSM	Recursive Subgraph Matching [127]
	DeA	De-Anonymization [64]
	ADA	Adaptive De-Anonymization [64]
	BDK	Backstrom et al.'s attacks [27]
	NS	Narayanan-Shmatikov's attack [104]
	NSR	Narayanan et al.'s attack [102]
	NKA	Nilizadeh et al.'s attack [108]
	PFG	Pedarsani et al.'s attack [112]
	YG	Yartseva-Grossglauser's attack [151]
	KL	Korula-Lattanzi's attack [69]
	JLSB	Ji et al.'s attack [63]

for high-level data mining tasks and applications, e.g., sense-making, search for similar users, user classification, reliable email, influence maximization. On the other hand and more seriously, to the best of our knowledge, no work (including existing DA works) actually evaluated the resistance of state-of-the-art anonymization techniques against modern SDA attacks.

To address these concerns, we comprehensively analyze the utility of existing graph data anonymization algorithms in this section and defer the detailed resistance analysis to Section 7.4. Before performing the analysis, we first present the used utility metrics, which can be classified as *graph utility metrics* or *application utility metrics*.

7.2.1 Graph Utility Metrics

Graph utility captures how the anonymized data preserves fundamental structural properties of the original graph after applying an anonymization technique. Particularly, we examine 12 graph utility metrics of existing anonymization schemes as follows[1].

- *Degree (Deg.)*, which refers to the degree distribution;
- *Joint Degree (JD)*, which refers to the degree distribution $\{p_{(x,y)}|p_{(x,y)}$ is the fraction of edges in a graph that connect users of degree x and degree $y\}$;
- *Effective Diameter (ED)*, which is defined as the minimum number of hops in which 90% of all connected pairs of nodes can reach each other;
- *Path Length (PL)*, which refers to the distribution of the shortest path lengths between all pairs of users;
- *Local Clustering Coefficient (LCC) and Global Clustering Coefficient (GCC).* *Clustering coefficient* measures the degree to which users in graph data tend to

[1]Without of causing confusion, we interchangeably use node and user in this chapter.

cluster together. The LCC of a user quantifies how close its neighbors are to being a *clique*. For $i \in V$, it is defined as

$$\mathrm{LCC}_i = \frac{2|E_{N_i}|}{d_i(d_i - 1)}. \tag{218}$$

The GCC is based on triplets of users. Let n_t and n_c be the number of *triangles* and the number of *connected triples of users* in a graph, respectively. Then, GCC is defined as

$$\mathrm{GCC} = \frac{3n_t}{n_c}. \tag{219}$$

- *Closeness Centrality (CC)*, which is defined as the *inverse of the farness* of a user within a graph and measures how long it takes to spread information from a user to all other users sequentially;

- *Betweenness Centrality (BC)*, which quantifies the number of times a user acts as a bridge along the shortest path between two other users;

- *Eigen Vector (EV)*. The EV of the adjacency matrix A of a graph G is a non-zero vector $\mathbf{v}$ such that $A\mathbf{v} = \lambda\mathbf{v}$, where λ is some scalar multiplier;

- *Network Constraint (NC)*, which measures the extent to which a user links to others that are already linked to each other;

- *Network Resilience (NR)* [26], which measures how robust a graph is and is defined as the number of users in the *largest connected component* when users are removed from the graph in the degree decreasing order;

- *Infectiousness (Infe.)* [138], which measures the number of users infected by a disease, given that a randomly chosen user is infected and each infected user transmits this disease to its neighbors with some *infection rate*;

7.2.2 Application Utility Metrics

In reality, most data is published/shared for data/network mining tasks, high-level applications, etc. Therefore, besides examining data's fundamental structural utility, it is also crucial to ensure that the anonymized data is useful for practical applications. Toward this objective, we evaluate 7 popular application utility metrics for anonymization schemes as follows.

- *Role eXtraction (RX) [59].* Based on users' structural behavior, users in a graph can be labeled as having different roles, e.g., *clique members*, *periphery-nodes*. RX is an important operation for graph data that is useful for many network mining tasks such as sense-making. We measure the RX utility of an anonymization scheme using the method in [59].

- *Reliable Email (RE) [47].* RE is a whitelisting system leveraging users' neighborhoods to filter and block spam emails. To evaluate the structural utility of an anonymization scheme with respect to RE, we take a similar method as in [122] to compute the number of users who can be spammed by a fixed number of compromised neighbors in a graph.

- *Influence Maximization (IM) [52].* The IM problem seeks to find a set of θ users such that these θ users have the maximum influence to the network under some influence propagation model. IM is important for many real world applications, e.g., advertisements. For our purpose, we evaluate the IM application utility of an anonymization scheme using the recently proposed method in [52].

- *Minimum-sized Influential Node Set (MINS) [56].* MINS is another popular and important application utility metric that leverages a graph's structure to identify the minimum-sized set of influential nodes, such that all other nodes

in the network could be influenced with a probability above a threshold. MINS can be used in many meaningful applications, e.g., social problems alleviation, new products promotion. We evaluate the MINS application utility of an anonymization scheme using the recent method in [56].

- *Community Detection (CD) [150].* CD is a popular application on graph data which enables comprehensive analysis of a network structure and supports other applications, e.g., classification, routing (information propagation). To measure the CD utility of an anonymization scheme, we employ the hierarchical agglomeration algorithm proposed in [150].

- *Secure Routing (SR) [92].* The structure of graph data can also be used to improve the performance of secure routing for systems such as P2P systems. For our purpose, we evaluate the SR application utility of an anonymization scheme using the method designed in [92].

- *Sybil Detection (SD) [154].* In a Sybil attack, an adversary tries to subvert a system by forging multiple identities. Sybil attacks are a serious threat to both centralized and distributed systems, e.g., recommendation systems, anonymity systems. Recently, several effective schemes, e.g., SybilLimit [154], have been proposed to defend against Sybil attacks. For our purpose, we evaluate the SD application utility of an anonymization scheme using the method in [154].

7.2.3 Anonymization vs Utility

We are ready to analyze the utility performance of existing graph data anonymization techniques. We summarize the graph and application utilities, and Resistance to SDA attacks (R2SDA) (e.g., [63, 64, 104, 151]) of existing graph anonymization schemes in Table 13. We analyze the results in Table 13 as follows.

Table 13: Analysis of existing graph anonymization techniques. ✓ = preserving the utility, ◐= partially preserving the utility, ⧫ = conditionally preserving the utility depending on parameters and considered data (based on our analysis, it is necessary to distinguish *partially* and *conditionally* preserving a data utility. For instance, k-DA conditionally preserves the Deg. utility depending on k while *Add/Del* can partially preserve the Deg. utility for an arbitrary k), ✗ = not preserving the utility, and **n/a** = evaluation not available in existing works.

	graph utility												application utility							R2SDA
	Deg.	JD	ED	PL	LCC	GCC	CC	BC	EV	NC	NR	Infe.	RX	RE	IM	MINS	CD	SR	SD	
Naive	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗
Add/Del [152]	◐	⧫	◐	◐	⧫	⧫	◐	◐	◐	⧫	◐	◐	✗	◐	◐	⧫	✗	⧫	⧫	✗
Switch [152]	✓	⧫	⧫	◐	⧫	⧫	◐	◐	◐	◐	⧫	◐	⧫	◐	◐	⧫	⧫	◐	◐	✗
k-NA [158]	⧫	⧫	⧫	⧫	⧫	⧫	◐	⧫	⧫	◐	◐	◐	✗	◐	⧫	⧫	⧫	⧫	⧫	**n/a**
k-DA [88]	⧫	⧫	⧫	⧫	⧫	⧫	◐	⧫	⧫	◐	◐	◐	✗	◐	⧫	⧫	⧫	⧫	⧫	**n/a**
k-auto [160]	⧫	⧫	⧫	⧫	⧫	⧫	◐	⧫	⧫	◐	◐	◐	✗	◐	⧫	⧫	⧫	⧫	⧫	**n/a**
k-iso [38]	⧫	⧫	✗	✗	⧫	✗	✗	✗	✗	⧫	✗	✗	✗	✗	✗	⧫	⧫	✗	⧫	**n/a**
Aggregation [53]	⧫	⧫	⧫	⧫	⧫	⧫	◐	⧫	⧫	◐	◐	◐	✗	◐	⧫	⧫	⧫	⧫	⧫	**n/a**
Cluster [131]	⧫	⧫	⧫	⧫	⧫	⧫	◐	⧫	⧫	◐	◐	◐	✗	◐	⧫	⧫	⧫	⧫	⧫	**n/a**
DP [122]	⧫	⧫	⧫	◐	⧫	⧫	◐	⧫	⧫	⧫	⧫	◐	✗	◐	⧫	⧫	✗	⧫	⧫	**n/a**
DP [117, 118]	⧫	⧫	⧫	◐	⧫	⧫	◐	⧫	⧫	⧫	⧫	◐	✗	◐	⧫	⧫	✗	⧫	⧫	**n/a**
DP [137]	⧫	⧫	⧫	◐	⧫	⧫	◐	⧫	⧫	⧫	⧫	◐	✗	◐	⧫	⧫	✗	⧫	⧫	**n/a**
DP [142]	⧫	⧫	⧫	◐	⧫	⧫	◐	⧫	⧫	⧫	⧫	◐	✗	◐	⧫	⧫	✗	⧫	⧫	**n/a**
RW [97]	✓	⧫	⧫	⧫	⧫	✗	◐	⧫	⧫	◐	⧫	◐	✗	◐	⧫	✗	✗	◐	◐	**n/a**

(a) For the Naive ID removal scheme, it is straightforward that it preserves all the data utility. However, it is also the most vulnerable scheme to SDA attacks.

(b) *Add/Del* and *Switch* are both designed to protect the node and link privacy of graph data [152]. Since *Add/Del* randomly adds and deletes edges, which is an global edge edition operation and thus it may change many fundamental structural properties of a graph. It follows that it can conditionally or partially preserve both graph and application utilities. However, utilities like JD, GCC, NC, CD, and MINS would be destroyed if too many existing edges are deleted while new edges are added. For *Switch*, it switches two randomly selected qualified edges, which preserves the degree of each user. Consequently, *Switch* can preserve Deg. and partially preserve most other utilities. Furthermore, compared to *Add/Del*, *Switch* can conditionally preserve the RX and CD utilities which are destroyed in *Add/Del*. This is because that *Add/Del* randomly changes users' degree in the global edge edition process and thus some global structure-sensitive application utility is lost or significantly affected. Furthermore, *Add/Del* and *Switch* cannot defend against modern SDA attacks as shown in [63, 104, 127].

(c) The k-anonymity based anonymization schemes k-NA [158], k-DA [88], and k-auto [160] can partially/conditionally preserve the graph and most application utilities except for the RX utility. This is because the fundamental idea of k-anonymity based schemes is to make k users/subgraphs structurally similar. Therefore, there is a tradeoff between anonymity and utility. If k is large, more users will be structurally similar while more utility will be lost. On the other hand, if k is chosen to be small, more utility will be preserved at the cost of lower anonymity guarantee. Furthermore, since every user is guaranteed to be structurally similar to at least $k-1$ other users while the RX utility tries to distinguish users based on their structural differences, it turns out k-anonymity based schemes cannot preserve the RX utility. As we discussed before, k-iso achieves structure anonymization by partitioning the original graph into

k isomorphic subgraphs. Therefore, several fundamental properties of a graph will be destroyed, e.g., connectivity. It follows that several important graph and application utilities are lost in k-iso, e.g., PL, GCC, NR, Infe., RX, RE, IM, and SR. Finally, compared with other schemes, k-NA, k-auto, and k-iso have higher computational complexities.

(d) Similar to k-anonymity based schemes, the cluster based schemes [53, 131] can conditionally/partially preserve graph and application utilities except for RX. This is because the fundamental idea of cluster based schemes is to make the users within a cluster structurally indistinguishable. Therefore, to what extent these schemes can preserve data utility depends on the cluster size setting. Again, since RX is achieved based on users' structural difference, this utility is not preserved in cluster based schemes.

(e) For DP based schemes (e.g., [122, 142]), their main objective is to protect link privacy by perturbing the edges of a graph. The fundamental idea of these schemes is to make an anonymized graph structurally similar to its neighboring graphs and thus an adversary cannot infer the existence of an edge. Therefore, they can conditionally/partially preserve most graph and application utilities. However, if a high level of privacy is guaranteed, many edges in the graph are changed. Furthermore, similar to *Add/Del*, the edge perturbation in DP also belongs to global edge edition. Therefore, the global structure-sensitive high-level application utilities, e.g., RX, MINS, and CD, are destroyed or significantly reduced in DP based schemes.

(f) In the RW based scheme [97], link privacy is achieved by replacing a random walk path with an edge, and thus this scheme, theoretically, will not change the degree distribution of the original data. It follows that several utilities, e.g., Deg., RX, SD, NR, Infe., can be preserved or partially preserved. However, some other global utilities, e.g. JD, GCC, are lost in the RW based scheme due to the significant change of the overall graph structure.

Table 14: Analysis of existing graph DA techniques. SF = seed-free, AGF = auxiliary graph-free, SemF = semantics-free, A/P = active/passive attack, Scal. = scalable, Prac. = practical, Rob. = robust to noise, ✓ = true, ◐= partially true, ⧫ = conditionally true, and ✗ = false.

	SF	AGF	SemF	A/P	Scal.	Prac.	Rob.
BDK [27]	✓	✓	✓	A, P	✗	◐	✗
NS [104]	✗	✗	✓	P	✓	✓	✓
NSR [102]	✗	✗	✓	P	✓	✓	✓
NKA [108]	⧫	✗	✓	P	⧫	⧫	⧫
DV [127]	✗	✗	✓	P	⧫	⧫	✓
RST [127]	✗	✗	✓	P	⧫	⧫	✓
RSM [127]	✗	✗	✓	P	⧫	⧫	✓
PFG [112]	✓	✗	✓	P	✓	⧫	⧫
YG [151]	✗	✗	✓	P	✓	⧫	✓
DeA [64]	✗	✗	✓	P	✓	✓	✓
ADA [64]	✗	✗	✓	P	✓	✓	✓
KL [69]	✗	✗	✓	P	✓	⧫	✓
JLSB [63]	✓	✗	✓	P	✓	✓	✓

(g) From Table 13, no existing work evaluates the resistance of state-of-the-art anonymization schemes against modern SDA attacks. Although most of the schemes have nice theoretical privacy guarantees, unfortunately, that privacy analysis cannot guarantee that they can defend against modern SDA attacks due to the improper model of the adversary's auxiliary information, problematic assumptions, etc. Therefore, aiming to address this open problem, we evaluate the effectiveness of existing graph data anonymization schemes against modern SDA attacks in Sections 7.4 and 7.5.

7.3 Graph De-anonymization

In this section, we analyze the performance of existing graph data DA algorithms. For convenience, in the rest of this chapter, we denote Backstrom et al.'s attacks [27] by BDK (the initials of the authors), Narayanan-Shmatikov's attack [104] by NS, Narayanan et al.'s attack [102] by NSR, Nilizadeh et al.'s attack [108] by NKA,

Srivatsa-Hicks' three attacks [127] by DV, RST, and RSM, respectively, Pedarsani et al.'s attack [112] by PFG, Yartseva-Grossglauser's attack [151] by YG, Ji et al.'s two attacks [64] by DeA and ADA, respectively, Korula-Lattanzi's attack [69] by KL, and Ji et al.'s attack [63] by JLSB. We show our analytical results in Table 14 and discuss the result as follows.

(a) Except for BDK, all the existing SDA attacks are passive attacks and require auxiliary graphs to perform the attack, i.e., they employ the structural similarity between the the anonymized graph and the auxiliary graph to break the anonymity. However, when we examine the anonymization schemes in Table 13, we find that none properly consider such auxiliary information in their threat models.

(b) To perform BDK attacks [27], an adversary either has to insert some Sybil users in the dataset before the actual anonymized data release, or has to be an internal user that knows its neighborhoods. In either case, such attacks can only de-anonymize some users but cannot de-anonymize users in large scale. Furthermore, the attacks cannot tolerate any topological change of the original data. Therefore, BDK attacks are not scalable or robust. These attacks require that an adversary successfully launches Sybil users or be an internal user that obtains his neighborhoods.

(c) All the examined DA attacks are semantics-free. This is because the structural information itself is sufficient to perfectly or partially de-anonymize graph users. Furthermore, compared to semantics information, structural information is widely available in large scale, resilient to noise, and easily computable [63, 104, 127]. Following this fact, all the attacks except for BDK are (conditionally) scalable, practical, and robust.

(d) Specifically, DV, RST, and RSM [127] are conditionally scalable and practical. This is because they are not computationally feasible when the number of seeds is large. PFG [112] is conditionally practical and robust. This is because it is very sensitive to the graph density of the anonymized data. Generally, this attack is

suitable for sparse graphs however it has a significant performance degradation as the graph density increases. YG [151] is conditionally practical because it is designed to de-anonymize users of degree no less than 4 in the anonymized data. In many real world graph datasets, the users with degree less than 4 could dominate or take a significant portion of graph data based on the statistics in [63]. The conditional practicability of KL [69] comes from its improper assumption that $\Theta(\iota \cdot n)$ ($\iota \in (0, 1]$ is a constant and n is the number of nodes in a graph) seeds are available, which is too strong to hold for real world DA attacks. Note that, the community-level DA of NKA [108] is scalable (with complexity of $O(n^2)$). However, the NKA [108] is conditionally scalable, practical, and robust. This is because, if the community-level DA of NKA [108] is employed to enhance DV, RST, RSM, YG, and/or KL, it is conditionally scalable, practical, and/or robust. NS [104], NSR [102], DeA, ADA, and JLSB [63, 64] adaptively perform DA employing several heuristics based on a graph's local and global structural characteristics. It follows that they are scalable, practical, and robust as long as similarity exists between anonymized graphs and auxiliary graphs.

(e) Both seed-based attacks (e.g., NS, DV) and seed-free attacks (e.g., PFG, JLSB) have advantages depending on the application scenarios. On one hand, seed-based attacks are more stable with respect to de-anonymizing arbitrary anonymized graphs. The reason is straightforward since seed knowledge provides more auxiliary information to an adversary. On the other hand, it is possible that in some scenarios seeds are not available, and thus seed-free attacks are more general. Furthermore, if there is some error in the seed seeking phase (which is possible in real world attacks), seed-based attacks will suffer performance de-gradation or will possibly fail.

(f) From Table 14, we see that BDK attacks can be defended against by state-of-the-art anonymization algorithms. This is because an implicit assumption in BDK attacks is that data publishers only anonymize the data by *naive ID removal*, i.e.,

Table 15: DA attacks vs anonymization techniques. Naive = naive ID removal, EE = EE based schemes [152], k-anony. = k-anonymity based schemes [38, 88, 158, 160], Cluster = cluster based schemes [53, 131], DP = DP based schemes [117, 118, 122, 137, 142], RW = the random walk based scheme [97], and ✓, ⧫, and ✗ = the anonymization scheme is vulnerable, conditionally vulnerable, and invulnerable (i.e., resistant) to the DA attack, respectively.

	Naive	EE	k-anony.	Cluster	DP	RW
BDK [27]	✓	✗	✗	✗	✗	✗
NS [104]	✓	✓	⧫	⧫	✓	✓
NSR [102]	✓	✓	⧫	⧫	✓	✓
NKA [108]	✓	⧫	⧫	⧫	✗	✗
DV [127]	✓	✓	⧫	⧫	✓	✓
RST [127]	✓	✓	⧫	⧫	✓	✓
RSM [127]	✓	✓	⧫	⧫	✓	✓
PFG [112]	✓	✓	⧫	⧫	✓	✓
YG [151]	✓	✓	⧫	⧫	✓	✓
DeA [64]	✓	✓	⧫	⧫	✓	✓
ADA [64]	✓	✓	⧫	⧫	✓	✓
KL [69]	✓	✓	⧫	⧫	✓	✓
JLSB [63]	✓	✓	⧫	⧫	✓	✓

no edge change (e.g., addition, deletion, switching) happens during the anonymization. Evidently, this assumption does not hold in any state-of-the-art anonymization schemes. However, *whether state-of-the-art anonymization schemes can defend against SDA attacks (e.g., [63, 64, 104, 151]) is still an open problem.* To fill this void, we conduct such analysis in the following section.

7.4 Anonymization vs DA Analysis

As we analyzed in Tables 13 and 14, understanding the vulnerability/resistance of state-of-the-art graph data anonymization schemes against modern SDA attacks is still an open problem. After carefully analyzing existing anonymization and DA techniques, we summarize the ***vulnerability*** of existing anonymization schemes in Table 15. We further experimentally validate our analysis in Section 7.5. Below, we analyze and discuss the results in Table 15.

(a) It has been shown in both academia and in practice that the naive ID removal anonymization cannot protect graph data's privacy. Therefore, naive anonymization is vulnerable to all the existing SDA attacks.

(b) As we analyzed before, all other state-of-the-art anonymization schemes (e.g., EE, k-anony., Cluster, DP, and RW) are resistant to BDK attacks. Again, this is because an assumption of BDK attacks is that data is anonymized by the naive ID removal technique.

(c) For EE based anonymization schemes ([152]), they are conditionally vulnerable to NKA [108] and vulnerable to all the other modern SDA attacks [63, 64, 104, 151]. This is because although EE can partially modify the structure of a graph, to preserve data utility, many structural properties, e.g., neighborhood, degree distribution, closeness/betweenness centrality distribution, and path length distribution, are generally preserved. Therefore, given an auxiliary graph consisting of the same or overlapping group of users with the anonymized graph, powerful DA heuristics can be designed based on these structural properties to break the privacy of EE based anonymization schemes. Furthermore, the availability of seed users make such heuristics more robust to the noise introduced by EE. For instance, NS breaks EE by employing degree and neighborhood similarity [104], DV, RST, and RSM break EE by employing path length and neighborhood similarity [127], DeA and ADA break EE by employing centrality similarity [64], etc. As we analyzed in Table 13, EE based anonymization schemes (e.g., *Add/Del*) may destroy graphs' community utility, and thus they are conditionally vulnerable to NKA [108].

(d) k-anonymity based anonymization schemes ([38, 88, 158, 160]) are conditionally vulnerable to modern SDA attacks [63, 64, 104, 151]. The reasons are as follows: k-anonymity is initially designed for traditional relational data, which makes a user semantically indistinguishable with $k-1$ other users. Unlike relational data, which are

structurally independent of each other, users in graph data have strong structural correlation in addition to semantic similarity. When researchers extended k-anonymity to graph data, they extended the concept of traditional semantics to graph data as different structural properties (e.g., degree, neighborhood, and subgraph), and designed schemes to make k users structurally indistinguishable with respect to some structural semantics, i.e., degree, neighborhood, subgraph, etc. However, even if users in graph data cannot be distinguished with respect to some structural semantics, e.g., degree, neighborhood, subgraph, they can be de-anonymized by other structural semantics, e.g., path length distribution, closeness centrality, betweenness centrality, or the combinations of several structural semantics. Theoretically, the only way to make users indistinguishable with respect to all structural semantics is to make a graph *completely connected* or *disconnected*, which also implies that all the data utility is destroyed. Therefore, as long as some data utility is preserved in the anonymized data, k-anonymity based schemes are vulnerable to modern SDA attacks. The degree of vulnerability depends on how much data utility is preserved.

(e) Cluster based schemes ([53, 131]) are also conditionally vulnerable to modern SDA attacks [63, 64, 104, 151]. The analysis is similar to that of k-anonymity. The fundamental idea of cluster based schemes is to cluster users first and then to make the users within a cluster indistinguishable with respect to neighborhoods. Again, even if users are indistinguishable by neighborhoods, they can be de-anonymized by other structural semantics or the combinations of other semantics, e.g., centralities scores, path length distribution. Consequently, cluster based schemes are vulnerable as long as some data utility, especially graph utilities, are preserved in the anonymized data, and the vulnerability depends on the amount of data utility preserved.

(f) DP and RW based schemes ([97, 117, 118, 122, 137, 142]) are vulnerable to modern SDA attacks except NKA [108]. The reasons are as follows: First, they are designed with the objective of protecting the link privacy of graph data and no

dedicated node privacy protection techniques are considered. Second, to protect link privacy, the edges are perturbed in DP based schemes and random walk paths are replaced by edges in the RW based scheme, both with a nice theoretical privacy guarantee. However, after the edge anonymization process, many data utilities, e.g., degree, path length distribution, are still preserved. This implies that, given an auxiliary graph, users are still de-anonymizable based on several structural semantics under DP and RW based schemes. Furthermore, as shown by Narayanan et al. in [102], link privacy can be breached after de-anonymizing the users in an anonymized graph (we also employ the same approach to break users' link privacy [20]). Again, as we analyzed in Table 13, since DP and RW based schemes cannot preserve data's community utility, they are resistant to NKA.

In summary, based on our analysis, state-of-the-art anonymization schemes are still vulnerable to modern DA attacks. The fundamental reasons are: first, existing anonymization schemes only ensure that graph data users are indistinguishable with respect to some structural semantics (properties). However, other structural semantics, especially global ones, and the combinations of multiple structural semantics can still enable effective DA of users; and second, as one of the main objectives, all the anonymization schemes try to preserve as much data utility as possible. However, data utility from the adversary's perspective is equivalent to structural information, which can be used along with an auxiliary graph for conducting powerful DA attacks.

7.5 SecGraph

As we found when discussing existing anonymization and DA techniques, they all have limitations when evaluating the techniques' performance. For instance, it is still an open problem to understand the resistance/vulnerability of state-of-the-art anonymization schemes against modern DA attacks. To address this open problem, we implement a Secure Graph data publishing/sharing (SecGraph) system.

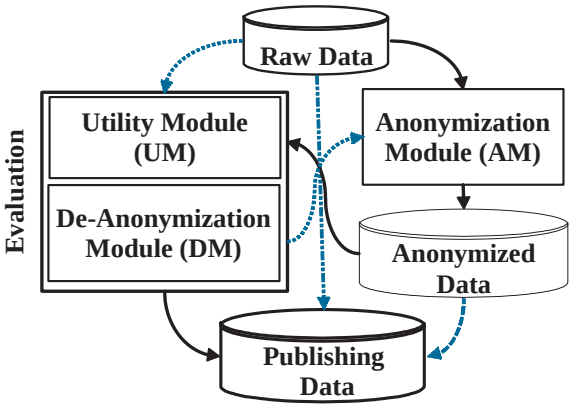


Figure 33: SecGraph: system overview.

7.5.1 System Overview

The overview of SecGraph is shown in Fig.33. SecGraph consists of three main modules: Anonymization Module (AM), Utility evaluation Module (UM), and DA evaluation Module (DM). The main functions of each module are briefly summarized as follows.

AM: the main function of this module is to anonymize raw graph data and generate anonymized data. In this module, we implement 11 state-of-the-art graph data anonymization schemes, including EE based algorithms [152], k-anonymity based algorithms and its variants [38, 88, 156, 158, 160], aggregation/class/cluster based algorithms [30, 53, 131], differential privacy based algorithms [117, 118, 122, 142], and the random walk based algorithm [97].

UM: in this module, we evaluate raw/anonymized data's utility with respect to the 12 graph utility metrics and 7 application utility metrics as defined in Section 7.2. With the UM, we can determine whether the data to be published/shared (e.g., the anonymized data) satisfies required utility requirements. We can also evaluate how an anonymization algorithm preserves data utility.

DM: in this module, we implement 15 SDA algorithms (all the existing SDA algorithms, to the best of our knowledge). By this module, the security of data to be

published/shared can be evaluated with real-world SDA attacks. More importantly, the effectiveness of an anonymization algorithm can be examined by this module, i.e., whether the anonymized data of an anonymization algorithm is resistant to modern SDA attacks.

We make further remarks on SecGraph and its modules and functions as follows.

(a) From Fig.33, raw data can be published/shared in multiple forms depending on the data owners' requirements on the security/privacy and utility of the data to be published. Each path in Fig.33 represents a data publishing scenario. For instance, the path *raw data* $\rightarrow$ *publishing data* means to publish the raw data directly. The path *raw data* $\rightarrow$ *AM* $\rightarrow$ *anonymized data* $\rightarrow$ *evaluation* $\rightarrow$ *publishing data* means that the raw data is anonymized first. Then, the anonymized data will be evaluated with respect to utility and/or practical de-anonymizability before actual publishing. The anonymization and evaluation process may be repeated several times until certain security and utility requirements are met.

(b) To the best of our knowledge, SecGraph is the first implemented uniform secure graph data publishing system, which systematically and comprehensively integrates state-of-the-art anonymization schemes, DA schemes, and graph/application utility measurements. The significance of SecGraph to the graph data anonymization and DA area lies in the following aspects. First, SecGraph enables data owners to conveniently and freely choose any modern anonymization algorithm to anonymize their data. They can also employ different evaluation modules to examine whether the anonymized data meets their security/privacy and utility requirements. Second, SecGraph is a uniform platform for testing and comparing different anonymization and DA algorithms. Previously, due to the lack of a uniform system, existing anonymization/DA algorithms are often proposed and implemented on separate platforms and

different environments/settings. Consequently, a number of implementation and evaluation differences (e.g., particular assumptions, models, evaluation datasets, programming, testing environments, parameter settings) limit researchers' understanding of the performance of existing anonymization and DA algorithms in different scenarios. However, as a uniform platform, SecGraph can reduce the evaluation bias caused by implementation and testing differences as much as possible. Therefore, SecGraph allows data owners to choose and compare the actual performance of different data anonymization algorithms on their data and thus to make the best decision. Additionally, SecGraph allows data anonymization researchers to compare their anonymization schemes to existing solutions as well as to examine their schemes' resistance against modern DA attacks. SecGraph also allows data DA researchers to evaluate the performance of new DA attacks by de-anonymizing the anonymized data of state-of-the-art anonymization schemes. Therefore, SecGraph is helpful to both data owners and researchers in conveniently applying existing schemes, comprehensively understanding existing algorithms, and effectively developing new anonymization/DA techniques.

(c) Besides providing a uniform platform, SecGraph is an easily portable and extendable system. First, the algorithms in SecGraph are implemented in Java and thus it is system independent. Second, all the modules of SecGraph are independent of each other, which means that each module can work individually. Additionally, as shown in Fig.33, multiple modules can also work together to perform data anonymization, utility evaluation, and de-anonymizability evaluation. Third, all the schemes/measurements within each module are independent, which means that they can be implemented, evaluated, and employed independently. Furthermore, newly developed anonymization/DA schemes and utility metrics can be easily integrated into SecGraph.

7.5.2 System Implementation

The implementation of SecGraph is as follows.

(a) In the AM, we implement 11 algorithms, which cover all the categories of state-of-the-art anonymization techniques. Specifically, the implemented anonyzation algorithms are naive ID removal, two EE based algorithms *Add/Del* [152] and *Switch* [152], two k-anonymity based algorithms k-DA [88] and k-iso [38], two cluster based algorithms bounded t-means clustering [131] and union-split clustering [131], three DP based algorithms Sala et al.'s scheme [122], Proserpio et al.'s scheme [117, 118], and Xiao et al.'s scheme [142], and one RW based algorithm [97]. Note that, we do not implement all the algorithms discussed in Section 7.2 even though we cover all the categories. The implementation criteria includes representativeness, scalability, and practicality, which led us to implement the latest, scalable, and practical schemes.

(b) In the UM, we implemented the 12 graph utility metrics and 7 application utility metrics as discussed in Section 7.2.

(c) In the DM, we implement all the 15 SDA attacks discussed in Section 7.3. To the best of our knowledge, these are all of the existing SDA attacks.

7.5.3 SecGraph-based Analysis

7.5.3.1 Primary Datasets

The employed datasets for evaluation are *Enron*, an email network consisting of 36.7K users and .2M edges, and *Facebook*, a Facebook friendship network in the New Orleans area consisting of 63.7K users and .82M edges [61, 63].

Table 16: Utility analysis of anonymization techniques. k is the number of modified edges for *Switch*, and the anonymization parameter for k-DA and Cluster, ϵ is the anonymization parameter for DP, t is the random walk step for RW, m is the number of edges in the original graph, and $\mathbb{D}$ is the diameter of the original graph ($\mathbb{D} = 11$ for Enron and $\mathbb{D} = 6$ for Facebook).

Utility	**Enron**										**Facebook**									
	Switch (vs. k)		k-DA (vs. k)		Cluster (vs. k)		DP (vs. ϵ)		RW (vs. t)		*Switch* (vs. k)		k-DA (vs. k)		Cluster (vs. k)		DP (vs. ϵ)		RW (vs. t)	
	$.05m$	$.1m$	5	50	5	50	300	50	2	$\mathbb{D}$	$.05m$	$.1m$	5	50	5	50	300	50	2	$\mathbb{D}$
Deg.	1	1	.9988	.9166	.9990	.9934	.9617	.8616	.9871	.9964	1	1	.9990	.9595	.9998	.9981	.9932	.9716	.9958	.9959
JD	.8725	.8338	.8928	.4183	.8216	.7055	.8496	.7363	.6972	.6438	.9941	.9804	.9947	.7328	.9872	.9024	.9755	.8263	.9678	.9362
ED	.9881	.9617	1.080	.9561	1.04	1.02	1.03	.9627	1.02	.9025	.9161	.8328	.9350	1.015	.9957	.9956	.9414	.9313	.9285	.8376
PL	.9954	.9887	.9891	.8934	.9994	.9905	.9565	.9839	.9963	.9657	.9618	.9159	.9999	.9946	.9999	1	.9960	.9653	.9706	.8965
LCC	.9830	.9631	.9972	.9809	.9966	.9797	.9528	.8328	.6785	.5985	.9204	.8303	.9998	.9983	.9968	.9947	.9793	.9437	.6239	.5543
GCC	.8967	.8013	.9921	.9283	.9774	.9097	.7755	.4609	.3107	.5383	.5180	.2241	.9847	.9986	.9766	.9937	.9522	.8702	.2552	.0334
CC	.9986	.9965	.9985	.9955	.9999	.9947	.9759	.9666	.9885	.9994	1	.9999	1	1	1	1	1	.9998	1	.9998
BC	.9859	.9812	.9691	.9019	.9936	.9733	.8360	.7406	.9613	.9246	.9787	.9494	.9790	.9515	.9983	.9897	.9779	.9518	.9935	.9669
EV	.9991	.9977	.9910	.8998	.9947	.9720	.9232	.8653	.9717	.9204	.9881	.9556	.9981	.9626	.9999	.9996	.9977	.9911	.9891	.9480
NC	.9984	.9962	.9999	.9991	.9996	.9956	.9977	.9596	.9042	.9028	.9995	.9986	1	1	1	1	.9987	.9934	.9928	.9942
NR	.9968	.9917	.9988	.9599	.9998	.9962	.9782	.8591	.9313	.8695	.9990	.9990	.9990	.9990	.9990	.9990	.9990	.9990	.9990	.9990
Infe.	.9627	.9597	.9604	.9411	.9427	.9413	.9662	.9593	.9664	.9446	.9748	.9704	.9758	.9695	.9730	.9719	.9730	.9699	.9788	.9778
PR	.9980	.9962	.9848	.8934	.9997	.9974	.9801	.9000	.8925	.9942	.9866	.9825	.9878	.9610	.9900	.9907	.9875	.9691	.9869	.9810
HS	.9991	.9977	.9910	.8998	.9947	.9720	.9232	.8653	.9717	.9204	.9326	.8780	.9711	.9789	.9648	.9625	.9626	.9322	.9283	.8655
AS	.9991	.9977	.9910	.8998	.9947	.9720	.9232	.8653	.9717	.9204	.9920	.9656	.9946	.9498	.9978	.9986	.9970	.9965	.9943	.9594
RX	.6575	.6009	.4561	.3173	.4512	.3685	.4196	.4116	.2955	.2680	.3494	.2608	.2974	.3139	.3902	.4652	.3483	.3134	.3250	.2772
RE	.9997	.9997	.9999	.9954	.9999	.9996	.9994	.9985	.9994	.9990	.9999	.9997	1	.9999	1	1	1	.9996	.9999	.9997
MINS	.7578	.6486	.9639	.9026	.9898	.9297	.7292	.3272	.1815	.1645	.6085	.4419	.9426	.9251	.9240	.9184	.8483	.7768	.2480	.1893
CD	.6251	.5411	.8454	.5339	.6794	.6692	.5095	.1028	.2531	.0569	.3536	.1986	.5043	.5887	.8558	.8523	.5027	.3213	.2860	.1205

7.5.3.2 Anonymization vs Utility

In this subsection, we evaluate the utility performance of anonymization algorithms. Due to the space limitation, we do not show the evaluation results of all the implemented algorithms. Particularly, we demonstrate the results of *Switch* [152], k-DA [88], *union-split clustering* [131], the improved version of Sala et al.'s DP scheme [117, 118, 122], and RW [97] which represent all the categories of anonymization algorithms. The evaluation methodology is that we first anonymize the original graph by an algorithm, and then measure how each data utility is preserved in the anonymized graph compared to the original graph. Specifically, when measuring utilities Deg., JD, PL, LCC, CC, BC, NC, NR, Infe., RX, and RE, we measure the *cosine similarity* between their distributions in the anonymized and original graphs; when measuring ED, GCC, and EV, we measure their *ratios* between the anonymized and original graphs; and when measuring MINS and CD, we measure their *Jaccard similarity* in the anonymized and original graphs.

We demonstrate the results in Table 16. (more results are available in [20]). The criteria for anonymization parameters settings are: (i) we follow the same/similar settings as in the original works of these anonymization schemes; and (ii) many data utilities can be preserved after anonymization. For the three graph utilities IM, SR, and SD, we only test them on small graphs, and put the results in [20]. We analyze the results in Table 16 as follows.

(a) Generally, the evaluation results in Table 16 are consistent with our analysis in Table 13. Most anonymization algorithms can partially or conditionally preserve most graph and application utilities. Therefore, most of the anonymized data can be employed for graph analytics, data mining tasks, and graph applications.

(b) Among all the graph utilities, JD and GCC are the most sensitive utilities to a graph's structure change, and thus they are the easiest ones to be destroyed by the anonymization algorithms. This is because these two utilities are very sensitive

to edge changes. Even if the degree distribution of the anonymized data remains the same as the original data, the JD distribution and GCC may change significantly.

(c) Compared to application utility, existing anonymization algorithms are better at preserving graph utility. For instance, most algorithms lost the RX utility and CD utility. This is because most application utilities depend on several graph utilities, e.g., the role of a user in RX depends on that user's degree, CC, BC, community attributes, and other structural characteristics. Therefore, application utilities are more easily affected than graph utilities, i.e., application utilities are more sensitive to graph's structural changes.

(d) No anonymization scheme is optimal in preserving every data utility. For instance, *Switch* is better than k-DA on preserving Deg. and JD while it is worse than k-DA on preserving GCC and MINS, and DP is better than RW on preserving LCC and GCC while it is worse than RW on preserving Deg. Therefore, when choosing an anonymization algorithm, it is better to take into account the specific application. Furthermore, RW has the most utility loss, e.g., GCC, RX, MINS, and CD, which is also consistent with our analysis in Table 13. This is because that the graph's global structure is significantly changed in RW by replacing random walk paths with edges.

7.5.3.3 DA Evaluation

In this subsection, we evaluate the performance of modern DA attacks. As we analyzed before, BDK [27], RST [127], and RSM [127] are not scalable/practical; NSR [102] and DeA [64] are simplified versions of NS [104] and ADA [64], respectively; and NKA [108] actually depends on other attacks, e.g., NS. Therefore, here, we focus on evaluating the seven general, practical, and scalable DA attacks: NS [104], DV (we replace its seed identification phase with a scalable one) [127], PFG [112], YG [151], ADA [64], KL [69], and JLSB [63]. Furthermore, PFG and JLSB are seed-free and the other five attacks are seed-based.

Table 17: Performance of DA attacks. s is the probability of generating the auxiliary and anonyized graphs from the original graph. Each value, e.g., 0.1277, in the table indicates the ratio of successfully de-anonymized users.

s	**De-anonymize Enron**							**De-anonymize Facebook**						
	NS	DV	PFG	YG	ADA	KL	JLSB	NS	DV	PFG	YG	ADA	KL	JLSB
.60	.0037	.1277	.0739	.0310	.1305	.1596	.1191	.0018	.1563	.1087	.2832	.1568	.0599	.1473
.65	.0039	.1601	.0937	.0410	.1651	.1814	.1460	.0020	.1998	.1402	.3346	.2005	.0747	.1799
.70	.0054	.1969	.1397	.0725	.2013	.2026	.1723	.0031	.2437	.1523	.4124	.2444	.0841	.2094
.75	.0055	.2244	.1349	.1004	.2307	.2152	.1958	.8712	.3068	.2041	.4554	.3078	.1196	.2574
.80	.0061	.2841	.1837	.1014	.2896	.2519	.2474	.9056	.3802	.2586	.4970	.3805	.1508	.3042
.85	.3420	.3481	.2180	.1531	.3522	.3123	.2971	.9231	.4561	.3073	.5402	.4576	.1817	.3559
.90	.3660	.4004	.2736	.1885	.4043	.3389	.3443	.9414	.5659	.3977	.5737	.5670	.2552	.4289
.95	.3937	.5814	.4370	.2277	.5898	.5209	.5438	.9527	.7407	.5584	.6071	.7422	.3989	.5542

First, employing the same Enron and Facebook datasets as before, we evaluate the DA performance of the seven DA attacks. The evaluation methodology is generally the same as in previous works [63, 64, 104, 108, 112, 127, 151]: we first randomly sample two graphs with probability s from the original data as the anonymized graph and auxiliary graph respectively, and then employ the auxiliary graph to de-anonymize the anonymized graph. Furthermore, for seed-based attacks, e.g., NS, DV, YG, ADA, and KL, we feed them 50 pre-identified seed mappings. The DA performance of the evaluated attacks with respect to different s is shown in Table 17. From Table 17, we have the following observations.

(a) With the increase of s, more users can be successfully de-anonymized under each algorithm. For instance, DV successfully de-anonymizes 12.77% Enron users when $s = .6$ while 58.14% Enron users when $s = .95$. The reason is evident. Since a large s implies that the anonymized graph and the auxiliary graph are more structurally similar, more accurate structural information can be employed by all the SDA algorithms. Hence, better DA performance can be achieved.

(b) Generally, all the algorithms have their advantages in some specific scenarios, and no algorithm is the best in all the cases. For instance, to de-anonymize Enron, KL has the best performance when $s = .6$ while ADA has the best performance when $s = .95$. Similarly, to de-anonymize Facebook, YG has the best performance when $s = .6$ while NS has the best performance when $s = .95$. Multiple reasons are responsible for the results such as the similarity between the anonymized and auxiliary graphs, the density of the anonymized/auxiliary graph, the heuristics employed by an algorithm, etc.

(c) According to the results, NS is more suitable for the scenarios where the anonymized and auxiliary graphs are highly similar while unsuitable when they are not sufficiently similar, e.g., it can successfully de-anonymize 95.27% Facebook users

when $s = .95$ while only 0.18% users when $s = .6$. The reason is because NS mainly employs local graph structural properties to adaptively conduct user DA, and thus is sensitive to users' local structural characteristics. When s is small, most users are indistinguishable with respect to their local structures, e.g., degree, followed by poor DA performance.

(d) Compared to NS, the other attacks, especially DV, PFG, ADA, and JLSB, are more stable even with a small s. For instance, when $s = .6$, DV, PFG, ADA, and JLSB can successfully de-anonymize 15.63%, 10.87%, 15.68%, and 14.73% Facebook users, respectively. This is because these attacks mainly employ global graph characteristics (e.g., clossness centrality, the distance vector to seeds) to perform the DA, which are more resilient to noise.

(e) For the seed-free attacks, PFG and JLSB, they can achieve comparable performance as seed-based attacks in most scenarios even without any seed information. For instance, when $s = .95$, PFG and JLSB can de-anonymize 43.7% and 54.38% Enron users, respectively, which are better than several seed-based algorithms and further demonstrate the power of structure-based attacks. The reason for the effectiveness of seed-free attacks is that in most cases, the combination of a user's local and global structural characteristics, e.g., degree, neighborhood degree distribution, closeness/betweenness centrality, is sufficient to distinguish him/her from other users.

7.5.3.4 Robustness of Modern SDA Attacks

The robustness of modern DA attacks with respect to graph noise (e.g., adding fake edges and deleting true edges) has been extensively evaluated in existing works [63, 64, 104, 127]. However, to the best our knowledge, no existing work has evaluated the robustness of any seed-based de-anonymizaton attack to incorrect seed mappings. Employing Enron and Facebook, we address this open issue by conducting such an

Table 18: DA robustness with respect to seed errors. Each algorithm is provided with 50 seed mappings, and Λ_e/Λ indicates the percentages of incorrect seed mappings. Each value in the table indicates the ratio of successfully de-anonymized users.

$\frac{\Lambda_e}{\Lambda}$	**De-anonymize Enron**					**De-anonymize Facebook**				
	NS	DV	YG	ADA	KL	NS	DV	YG	ADA	KL
4%	.341	.342	.148	.336	.302	.922	.456	.537	.442	.183
6%	.341	.342	.133	.329	.303	.917	.456	.528	.440	.183
8%	.338	.348	.135	.329	.310	.918	.456	.542	.428	.184
10%	.007	.348	.147	.323	.310	.918	.456	.536	.420	.182
12%	.007	.348	.142	.313	.311	.915	.456	.529	.414	.185
14%	.006	.348	.112	.306	.307	.916	.456	.526	.403	.186
16%	.006	.348	.129	.297	.303	.916	.456	.525	.394	.184
18%	.006	.348	.099	.293	.308	.913	.456	.533	.380	.183
20%	.006	.348	.126	.285	.306	.913	.456	.518	.356	.179
22%	.005	.348	.125	.280	.303	.912	.456	.531	.347	.182
24%	.005	.348	.116	.268	.304	.910	.456	.521	.332	.180
26%	.005	.348	.118	.255	.303	.889	.456	.528	.319	.179
28%	.004	.348	.112	.253	.300	.886	.456	.520	.309	.182
30%	.004	.348	.120	.247	.307	.884	.456	.522	.283	.180
32%	.004	.348	.106	.235	.305	.888	.456	.521	.270	.178
34%	.004	.348	.081	.230	.304	.887	.456	.521	.259	.178
36%	.004	.348	.084	.216	.300	.889	.456	.505	.245	.182
38%	.004	.347	.096	.199	.301	.888	.456	.493	.230	.178
40%	.004	.347	.065	.186	.302	.886	.456	.505	.214	.179
42%	.003	.347	.071	.182	.302	.882	.456	.516	.195	.181
44%	.003	.347	.106	.169	.303	.881	.456	.495	.185	.180
46%	.003	.347	.050	.160	.299	.881	.456	.480	.173	.177
48%	.003	.347	.059	.153	.297	.881	.456	.497	.161	.180
50%	.002	.347	.063	.146	.298	.874	.456	.475	.148	.176

evaluation and the results are shown in Table 18. We analyze the results in Table 18 as follows.

(a) Generally, all the DA algorithms are robust with respect to incorrect seed mappings in most scenarios. This is because during the DA process, most algorithms also employ other seed-independent structural properties, e.g., degree, closeness/betweenness centrality, in addition to relying on seed-dependent structural properties. Even for the pure seed-based DA attacks, e.g., YG and KL, they perform DA in the decreasing order of user degrees. Therefore, the negative impacts of incorrect seed mappings can be partially offset, i.e., even with some incorrect seed mappings, many users are still distinguishable with respect to their structural characteristics.

(b) For all algorithms, when incorrect seed mappings increase, fewer users can be correctly de-anonymized. The reason is evident: more incorrect seed mappings imply more incorrect seed-dependent structural information is provided to each algorithm, followed by the degradation of the DA performance of each algorithm.

(c) When de-anonymizing Enron, the performance of NS has a significant drop when the percentage of incorrect seed mappings is increased from 8% to 10%. This is because of the seed transitional phenomena as observed in [104], i.e., when the correct effective seed-dependent structural information is below/above some crucial threshold, NS's performance has a significant transition.

(d) DV is much more stable than other algorithms. This is because it is a pure global structure-based attack and thus incorrect seed mappings have minimum impact on it.

7.5.3.5 Anonymization vs DA

Now, we evaluate the effectiveness of state-of-the-art anonymization techniques against modern DA attacks employing Enron and Facebook. The methodology is that

Table 19: Anonymization vs DA. The seed-based algorithms are provided with 50 seeds and the anonymization parameters are chosen according to the same criteria as in Table 16.

	s	**Enron**										**Facebook**									
		Switch (k)		k-DA (k)		Cluster (k)		DP (ϵ)		RW (t)		*Switch* (k)		k-DA (k)		Cluster (k)		DP (ϵ)		RW (t)	
		5	10	5	50	5	50	300	50	2	$\mathbb{ID}$	5	10	5	50	5	50	300	50	2	$\mathbb{ID}$
NS	.85	.0072	.0052	.3702	.0088	.3722	.3707	.0091	.0055	.0015	.0015	.8973	.8247	.9454	.9402	.9456	.9442	.9317	.8914	.0008	.0006
	.90	.0077	.0054	.3822	.0105	.3900	.3839	.0095	.0060	.0015	.0015	.9063	.8427	.9520	.9495	.9519	.9508	.9393	.8944	.0008	.0007
	.95	.3577	.0064	.4033	.0418	.4049	.4064	.3946	.0064	.0015	.0016	.9162	.8583	.9570	.9559	.9569	.9558	.9453	.9130	.0000	.0007
DV	.85	.1261	.0813	.1433	.0437	.2120	.1408	.1160	.0701	.1923	.2412	.1716	.0926	.2411	.0588	.3340	.3368	.2324	.0736	.1530	.1271
	.90	.1546	.0956	.1765	.0517	.2564	.1637	.1394	.0733	.2129	.2169	.2124	.1147	.2999	.0758	.4113	.4090	.3623	.0802	.1604	.1322
	.95	.2121	.1366	.2548	.0753	.3745	.2215	.1821	.0858	.2072	.2190	.3006	.1586	.4210	.1161	.5767	.5656	.4087	.1016	.1591	.1332
PFG	.85	.0667	.0422	.0692	.0214	.1116	.0683	.0489	.0365	.1578	.2131	.0706	.0395	.0703	.0154	.1191	.1155	.0891	.0206	.1349	.1190
	.90	.0805	.0478	.0810	.0263	.1317	.0789	.0571	.0390	.1711	.2012	.0978	.0497	.0946	.0213	.1480	.1595	.1870	.0223	.1382	.1217
	.95	.1193	.0695	.1123	.0353	.1978	.0952	.0755	.0479	.1714	.2074	.1378	.0725	.1317	.0332	.2034	.2330	.1756	.0295	.1397	.1216
YG	.85	.1373	.0969	.1646	.0289	.1576	.1570	.1549	.0664	.0394	.0323	.5437	.5056	.5816	.5086	.5897	.5805	.5404	.4347	.0356	.0210
	.90	.1716	.1037	.1612	.0253	.1868	.1710	.1577	.0736	.0404	.0342	.5681	.5182	.6089	.5129	.6036	.5980	.5702	.4818	.0372	.0222
	.95	.1730	.1197	.2155	.3785	.1971	.2064	.1884	.0838	.0418	.0348	.5821	.5439	.6208	.5504	.6223	.6190	.5716	.4538	.0346	.0231
ADA	.85	.1262	.0820	.1468	.0445	.2130	.1418	.1160	.0701	.0771	.0731	.1724	.0926	.2425	.0603	.3358	.3379	.2337	.0749	.0985	.0725
	.90	.1543	.0964	.1795	.0534	.2588	.1652	.1394	.0729	.0855	.0704	.2129	.1146	.3026	.0776	.4124	.4103	.3639	.0823	.1008	.0764
	.95	.2139	.1381	.2605	.0768	.3777	.2230	.1823	.0855	.0872	.0733	.3019	.1589	.4245	.1186	.5780	.5667	.4105	.1038	.1041	.0784
KL	.85	.0904	.0811	.0997	.0357	.0965	.0689	.0745	.0331	.0900	.0729	.0799	.0764	.0819	.0683	.0788	.0762	.0769	.0313	.1099	.0737
	.90	.1077	.0970	.1202	.0549	.1134	.0918	.0874	.0319	.0939	.0744	.0979	.0939	.1013	.0848	.0960	.0863	.1249	.0317	.1099	.0715
	.95	.1381	.1150	.1936	.0978	.2052	.1686	.1719	.0376	.0994	.0776	.1350	.1331	.1418	.1265	.1294	.1206	.1450	.0600	.1171	.0754
JLSB	.85	.0692	.0440	.0798	.0234	.1248	.0854	.0886	.0656	.0709	.0720	.1453	.0786	.2025	.0595	.2618	.2673	.1958	.0768	.0901	.0681
	.90	.0886	.0536	.1046	.0296	.1618	.1135	.1070	.0664	.0767	.0728	.1673	.0911	.2335	.0708	.3001	.3094	.3050	.0777	.0911	.0699
	.95	.1846	.1189	.2381	.0746	.3317	.2319	.1449	.0814	.0838	.0740	.2180	.1174	.3111	.1096	.3983	.3924	.3142	.0950	.0940	.0734

we first employ different anonymization techniques to anonymize Enron/Facebook. Then, we sample an auxiliary graph from Enron/Facebook with probability s. Finally, we employ different DA algorithms to de-anonymize the anonymized data using the auxiliary graph. We show the results in Table 19 and analyze the results as follows.

(a) All the state-of-the-art graph anonymization algorithms are vulnerable to some or all of the modern SDA attacks, which confirmed our analytical results in Table 15. For instance, when $s = .85$, NS can still successfully de-anonymize more than 80% Facebook users anonymized by *Switch*, k-DA, Cluster, or DP, and DV can successfully de-anonymize 15.3% Facebook users anonymized by RW ($t = 2$). Similarly, when $s = .85$, NS can successfully de-anonymize more than 35% Enron users anonymized by k-DA ($k = 5$), Cluster ($k = 5, 50$), YG can successfully de-anonymize 13.73% and 15.49% Enron users anonymized by *Switch* ($k = .05m$) and DP ($\epsilon = 300$) respectively, and DV can successfully de-anonymize 19.23%/24.12% Enron users anonymized by RW with $t = 2/11$. Based on the results, we conclude that modern SDA attacks are very powerful. As we analyzed in Table 15, two fundamental reasons make state-of-the-art graph anonymization algorithms vulnerable. First, in existing graph anonymization schemes, graph users are only indistinguishable with respect to some structural properties/semantics. However, several other structural properties or the combinations of them can still enable effective graph user DA. Furthermore, the design philosophy of existing anonymization schemes is to preserve as much data utility as possible. However, data utility can be used to conduct powerful SDA attacks. Therefore, it is still an open problem to design effective graph data anonymization algorithms which can defend against modern SDA attacks.

(b) Generally, when s is large and the anonymization level (e.g., k for *Switch* and k-DA) is low, more users can be correctly de-anonymized. The reason is straightforward. A large s implies more structural information of the original graph can be preserved in

the auxiliary graph and thus more accurate structural characteristics can be employed for DA. Meanwhile, a low anonymization level implies less perturbation applied to the original graph's structure followed by the anonymized graph is more structurally similar to the original graph and thus is easier to be de-anonymized.

(c) Among all the DA attacks, NS, YG, and ADA perform better than other attacks in most scenarios. This is because they mainly employ the combinations of several local structural characteristics to conduct the DA. According to our utility analysis in Table 13 and evaluation results in Table 16, most existing anonymization algorithms can preserve most graph utilities, especially the local graph utilities, e.g., Deg., LCC. It turns out that the graph utility preserved by anonymization algorithms can be used by DA attacks to conduct effective DA. Therefore, in the scenarios where an anonymization algorithm preserves more data utility, the corresponding dataset is more vulnerable to modern SDA attacks.

(d) Among all the anonymization techniques, RW has better performance than others in most of the cases. The reason is that, a random walk path of length t is replaced by an edge in RW. It follows that the original graph structure is significantly changed. Therefore, a RW-anonymized graph is more resistant to DA attacks. However, RW achieves such DA resistance at the cost of sacrificing more data utility compared with other anonymization techniques, which is consistent with our utility analysis and evaluation results in Tables 13 and 16. Furthermore, we can also find that in most scenarios, existing anonymization techniques can degrade the performance of SDA attacks. Again, as shown in Tables 13 and 16, some data utilities are also degraded/lost.

7.6 Chapter Summarization

In this chapter, we propose, implement, and evaluate SecGraph (available at [20]), an *open-source* secure graph data publishing/sharing system. Within SecGraph, we

systematically analyze, implement, and evaluate 11 graph data anonymization algorithms, 19 data utility metrics, and 15 modern SDA attacks. To the best of our knowledge, SecGraph is the first such system that provides a *uniform platform* enabling data owners to anonymize and evaluate the security of their data, and simultaneously enabling researchers to conduct fair studies of existing or newly developed anonymization/DA techniques. Leveraging SecGraph, we conduct extensive experimental evaluations. The results demonstrate that (*i*) most anonymization schemes can partially or conditionally preserve most graph utility but lose some application utility; (*ii*) no DA attack is optimum in all scenarios. The actual DA performance depends on several factors; and (*iii*) all the state-of-the-art anonymization schemes are vulnerable to modern SDA attacks. Based on our findings and analysis, we discuss the future research directions and challenges of graph data anonymization and DA.

CHAPTER VIII

CONCLUSION

In this dissertation, we study the security of anonymized big graph data. Our main contributions include: *new De-Anonymization (DA) attacks*, *comprehensive anonymity, utility, and de-anonymizability quantifications*, and *a secure graph data publishing/sharing system SecGraph*.

New DA Attacks. We present two novel graph DA frameworks: *cold start single-phase Optimization-based DA* (ODA) and *De-anonymizing Social-Attribute Graphs* (De-SAG). Unlike existing seed-based DA attacks, ODA does not priori knowledge. In addition, ODA's DA results can facilitate existing DA attacks by providing more seed information. We validated ODA's performance leveraging real world graph datasets. Gowalla (196,591 users and 950,327 edges) and Google+ (4,692,671 users and 90,751,480 edges). The results demonstrate that about $77.7\% - 83.3\%$ of the users in Gowalla and $86.9\% - 95.5\%$ of the users in Google+ are de-anonymizable, which implies seed-free DA is implementable and powerful in practice. De-SAG takes into account both graph structure and attribute information. Through extensive evaluations leveraging real world graph data, we demonstrated that De-SAG can significantly enhance existing SDA attacks. For instance, when de-anonymizing a Facebook dataset (4,039 users, 88,234 user-user links, 1,283 attributes, 37,257 user-attribute links), De-SAG has a $3.82 \sim 10.1$ times better DA performance than state-of-the-art structure-based deanonymization attacks.

Comprehensive Graph Anonymity, Utility, and De-anonymizability Quantifications. We developed new techniques that enable comprehensive graph data anonymity, utility, and de-anonymizability evaluation. First, we proposed the first

seed-free graph de-anonymizability quantification framework under a general data model. In our quantification, we answered several fundamental open problems: why graph data can be de-anonymized based only on the topological information? what are the conditions for perfect and $(1-\epsilon)$-perfect seed-free DA, where ϵ is the error tolerated by a DA scheme? and what portion of users can be de-anonymized in a graph dataset? Thus, our quantification provides the theoretical foundation for seed-free SDA attacks.

Second, we conducted the first seed-based quantification on the perfect and partial de-anonymizability of graph data both under the Erdös-Rényi (ER) model and in general scenarios. Our quantification provides the mathematical foundation for existing seed-based SDAs and closes the gap between seed-based DA practice and theory.

Third, we conducted the first attribute-based anonymity analysis for Social-Attribute Graph (SAG) data under both preliminary and general data models. Our theoretical results demonstrate that the non-personal identifiable information can also lead to significant anonymity loss of graph data. Our attribute-based anonymity analysis together with existing structure-based de-anonymizability quantifications provide data owners and researchers a more complete understanding of the privacy of graph data.

Fourth, we conducted a comprehensive quantification of the correlation of graph anonymity, utility, and de-anonymity. This is the first work on quantifying the Anonymity-Utility-De-anonymity (AUD) correlation of graph data and providing close-forms to explicitly demonstrate such correlation.

Finally, based on our quantifications, we conducted large-scale evaluations leveraging 100+ real world graph datasets generated by various computer systems and services. Using the evaluations, we demonstrated the datasets' anonymity, utility, and de-anonymizability, as well as the significance and validity of our quantifications.

SecGraph: A Secure Graph Data Publishing/Sharing System. We designed, implemented, and evaluated a uniform and open-source Secure Graph data publishing/sharing (SecGraph) system (available at [20]). SecGraph enables data owners to anonymize their data using state-of-the-art anonymization techniques, measure the anonymized datas graph and application utilities, and comprehensively evaluate their datas actual vulnerability against modern DA attacks. To the best of our knowledge, SecGraph is the first such system publicly available to both academia and industry. More importantly, SecGraph provides the first uniform platform that enables researchers to conduct accurate comparative studies of anonymization/DA techniques, and to comprehensively understand the resistance/vulnerability of existing or newly developed anonymization techniques, the effectiveness of existing or newly developed DA attacks, and graph and application utilities of anonymized data.

Bibliography

[1] *http://socialcomputing.asu.edu/pages/datasets.*

[2] *http://www.informatik.uni-trier.de/∼ley/db/.*

[3] *http://arnetminer.org.*

[4] *http://www.sigkdd.org/kddcup/index.php.*

[5] *http://icwsm.org/2013/datasets/datasets/.*

[6] *https://blog.twitter.com/2014/introducing-twitter-data-grants.*

[7] *http://danzarrella.com/about-my-facebook-sharing-dataset-and-methodology#.*

[8] *https://www.kddcup2012.org/c/kddcup2012-track1.*

[9] *http://www.google.com/policies/privacy/.*

[10] *https://www.facebook.com/note.php?note_id=%20322194465300.*

[11] *https://twitter.com/privacy.*

[12] *http://www.andrew.cmu.edu/user/rkoganti/realistic.html.*

[13] *http://www.slideshare.net/jlcaut/ebola-hemoragic-fever-propagation-in-a-modern-city-using-sir-model.*

[14] *https://www.data.gov/.*

[15] *http://www.cs.berkeley.edu/ stevgong/dataset.html.*

[16] *http://snap.stanford.edu/data/index.html.*

[17] *http://www.casos.cs.cmu.edu/computational_tools/data2.php.*

[18] *http://crawdad.cs.dartmouth.edu/.*

[19] *http://networkdata.ics.uci.edu/resources.php.*

[20] *http://www.ece.gatech.edu/cap/secgraph/.*

[21] *Google programming contest*, 2002.

[22] *http://www.cpc.unc.edu/projects/addhealth*, 2008.

[23] Aggarwal, G., Feder, T., Kenthapadi, K., Khuller, S., Panigrahy, R., Thomas, D., and Zhu, A., "Achieving anonymity via clustering," *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS)*, pp. 153–162, 2006.

[24] Aggarwal, G., Feder, T., Kenthapadi, K., Motwani, R., Panigrahy, R., Thomas, D., and Zhu, A., "Approximation algorithms for k-anonymity," *Journal of Privacy Technology (JOPT)*, pp. 1–18, 2005.

[25] Albert, R., Jeong, H., and Barabasi, A.-L., "Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, pp. 130–131, 1999.

[26] Albert, R., Jeong, H., and Barabasi, A.-L., "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[27] Backstrom, L., Dwork, C., and Kleinberg, J., "Wherefore art thou r3579x? anonymized social networks, hidden patterns, and structural steganography," *Proceedings of the 16th international conference on World Wide Web*, pp. 181–190, 2007.

[28] Bearman, P., Moody, J., and Stovel, K., "Chains of affection: The structure of adolescent romantic and sexual networks," *American Journal of Sociology*, vol. 110, no. 1, pp. 44–91, 2004.

[29] Beckett, L., "Everything we know about what data brokers know about you," *http://www.propublica.org/article/everything-we-know-about-what-data-brokers-know-about-you*, 2015.

[30] Bhagat, S., Cormode, G., Krishnamurthy, B., and Srivastava, D., "Class-based graph anonymization for social network data," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 766–777, 2009.

[31] Bollobás, B., "Random graphs (second edition)," *Cambridge University Press*, 2001.

[32] Brickell, J. and Shmatikov, V., "The cost of privacy: Destruction of data-mining utility in anonymized data publishing," *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 70–78, 2008.

[33] Cao, J., Karras, P., Raïssi, C., and Tan, K., "ρ-uncertainty: Inference-proof transaction anonymization," *Proceedings of the VLDB Endowment*, vol. 3, no. 1, pp. 1033–1044, 2010.

[34] Chang, C., Thompson, B., Wang, H., and Yao, D., "Towards publishing recommendation data with predictive anonymization," *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security (ASIACCS)*, pp. 24–35, 2010.

[35] Chen, R., Acs, G., and Castelluccia, C., "Differentially private sequential data publication via variable-length n-grams," *Proceedings of the 2012 ACM conference on Computer and communications security (CCS)*, pp. 638–649, 2012.

[36] Chen, R., Fung, B., Desai, B., and Sossou, N., "Differentially private transit data publication: A case study on the montreal transportation system," *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 213–221, 2012.

[37] Chen, R., Mohammed, N., Fung, B., Desai, B., and Xiong, L., "Publishing set-valued data via differential privacy," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 1087–1098, 2011.

[38] Cheng, J., Fu, A., and Liu, J., "K-isomorphism: Privacy preserving network publication against structural attacks," *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pp. 459–470, 2010.

[39] Choromanski, K., Jebara, T., and Tang, K., "Adaptive anonymity via b-matching," *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, pp. 3192–3200, 2013.

[40] Clayton, M., "Nsa data-mining 101: two 'top secret' programs and what they do," *http://www.csmonitor.com/USA/2013/0607/NSA-data-mining-101-two-top-secret-programs-and-what-they-do*.

[41] Cormode, G., "Personal privacy vs population privacy: Learning to attack anonymization," *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 1253–1261, 2011.

[42] Cormode, G., Srivastava, D., Li, N., and Li, T., "Minimizing minimality and maximizing utility: Analyzing method-based attacks on anonymized data," *Proceedings of the VLDB Endowment*, vol. 3, no. 1, pp. 1045–1056, 2010.

[43] Data-Broker, "The data brokers: Selling your personal information," *http://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information/*, 2015.

[44] Dwork, C., "Differential privacy," *Automata, languages and programming (ICALP)*, pp. 1–12, 2006.

[45] Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M., "Our data, ourselves: Privacy via distributed noise generation," *Advances in Cryptology-EUROCRYPT 2006*, pp. 486–503, 2006.

[46] Facebook *https://www.facebook.com/*.

[47] Garriss, S., Kaminsky, M., Freedman, M. J., Karp, B., Mazières, D., and Yu, H., "Re: Reliable email," *Proceedings of the 3rd conference on Networked Systems Design & Implementation*, vol. 3, pp. 22–22, 2006.

[48] Gehrke, J., Ginsparg, P., and Kleinberg, J. M., "Overview of the 2003 kdd cup," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 149–151, 2003.

[49] Gong, N. Z., Xu, W., Huang, L., Mittal, P., Stefanov, E., Sekar, V., and D.Song, "Evolution of social-attribute networks: Measurements, modeling, and implications using google+," *Proceedings of the 2012 ACM conference on Internet measurement conference (IMC)*, pp. 131–144, 2012.

[50] Goodin, D. *http://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/.*

[51] Google+ *https://plus.google.com/.*

[52] Goyal, A., Bonchi, F., and Lakshmanan, L. V. S., "A data-based approach to social influence maximization," *Proceedings of the VLDB Endowment*, vol. 5, no. 1, pp. 73–84, 2011.

[53] Hay, M., Miklau, G., Jensen, D., Towsley, D., and Weis, P., "Resisting structural re-identification in anonymized social networks," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 102–114, 2008.

[54] Hay, M., Rastogi, V., Miklau, G., and Suciu, D., "Boosting the accuracy of differentially private histograms through consistency," *Proceedings of the VLDB Endowment (VLDB)*, 2006.

[55] Hayes, B., "Connecting the dots: Can the tools of graph theory and social-network studies unravel the next big plot?," *American Scientist*, vol. 94, no. 5, pp. 400–404, 2006.

[56] He, J., Ji, S., Beyah, R., and Cai, Z., "Minimum-sized influential node set selection for social networks under the independent cascade model," *Proceedings of the 15th ACM international symposium on Mobile ad hoc networking and computing (Mobihoc)*, pp. 93–102, 2014.

[57] He, Y. and Naughton, J. F., "Anonymization of set-valued data via top-down, local generation," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 934–945, 2009.

[58] Health, A., "Deductive disclosure," *http://www.cpc.unc.edu/projects/addhealth/data/dedisclosure*, 2008.

[59] Henderson, K., Gallagher, B., Eliassi-Rad, T., Tong, H., Basu, S., Akoglu, L., Koutra, D., Li, L., and Faloutsos, C., "Rolx: Structural role extraction & mining in large graphs," *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 1231–1239, 2012.

[60] J. Leskovec, J. K. and Faloutsos, C., "Graphs over time: Densification laws, shrinking diameters and possible explanations," *Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining (KDD)*, pp. 177–187, 2005.

[61] Ji, S., Li, W., Gong, N., Mittal, P., and Beyah, R., "On your social network de-anonymizablity: Quantification and large scale evaluation with seed knowledge," *Proceedings of the 2015 Network and Distributed System Security (NDSS) Symposium*, pp. 1–15, 2015.

[62] Ji, S., Li, W., Mittal, P., Hu, X., and Beyah, R., "Secgraph: A uniform and open-source evaluation system for graph data anonymization and de-anonymization," *Proceedings of the 24th USENIX Security Symposium*, pp. 303–318, 2015.

[63] Ji, S., Li, W., Srivatsa, M., and Beyah, R., "Structural data de-anonymization: Quantification, practice, and implications," *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1040–1053, 2014.

[64] Ji, S., Li, W., Srivatsa, M., He, J., and Beyah, R., "Structure based data de-anonymization of social networks and mobility traces," *Proceedings of the Information Security (ISC)*, pp. 237–254, 2014.

[65] Kellaris, G. and Papadopoulos, S., "Practical differential privacy via grouping and smoothing," *Proceedings of the VLDB Endowment (VLDB)*, vol. 6, no. 5, pp. 301–312, 2013.

[66] Kifer, D. and Gehrke, J., "Injecting utility into anonymized datasets," *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pp. 217–228, 2006.

[67] Klimmt, B. and Yang, Y., "Introducing the enron corpus," *CEAS*, 2004.

[68] Korolova, A., Motwani, R., Nabar, S., and Xu, Y., "Link privacy in social networks," *Proceedings of the 17th ACM conference on Information and knowledge management (CIKM)*, pp. 289–298, 2008.

[69] Korula, N. and Lattanzi, S., "An efficient reconciliation algorithm for social networks," *Proceedings of the VLDB Endowment*, vol. 7, no. 5, pp. 377–388, 2014.

[70] Koudas, N., Srivastava, D., Yu, T., and Zhang, Q., "Distribution-based microdata anonymization," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 958–969, 2009.

[71] Kurucz, M., Benczúr, A., Csalogány, K., and Lukács, L., "Spectral clustering in telephone call graphs," *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp. 82–91, 2007.

[72] Lakshmanan, L., Ng, R., and Ramesh, G., "To do or not to do: The dilemma of disclosing anonymized data," *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 61–72, 2005.

[73] Lambiotte, R., Blondel, V. D., Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., and Dooren, P. V., "Geographical dispersal of mobile communication networks," *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 21, pp. 5317–5325, 2008.

[74] Lee, E. K., Chen, C. H., Pietz, F., and Benecke, B., "Disease propagation analysis and mitigation strategies for effective mass dispensing," *AMIA annual symposium proceedings*, pp. 427–427, 2010.

[75] Lee, J. and Clifton, C., "Differential identifiability," *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 1041–1049, 2012.

[76] LeFevre, K., DeWitt, D., and Ramakrishnan, R., "Workload-aware anonymization," *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 277–286, 2006.

[77] LeFevre, K., DeWitt, D. J., and Ramakrishnan, R., "Incognito: Efficient full-domain k-anonymity," *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp. 49–60, 2005.

[78] Leskovec, J., Huttenlocher, D., and Kleinberg, J., "Predicting positive and negative links in online social networks," *Proceedings of the 19th international conference on World wide web (WWW)*, pp. 641–650, 2010.

[79] Leskovec, J., Huttenlocher, D., and Kleinberg, J., "Signed networks in social media," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 1361–1370, 2010.

[80] Leskovec, J., Kleinberg, J., and Faloutsos, C., "Graph evolution: Densification and shrinking diameters," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 2–2, 2007.

[81] Leskovec, J., Lang, K., Dasgupta, A., and Mahoney, M., "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

[82] Li, C., Hay, M., Miklau, G., and Wang, Y., "A data- and workload-aware algorithm for range queries under differential privacy," *Proceedings of the VLDB Endowment (VLDB)*, vol. 7, no. 5, pp. 341–352, 2014.

[83] Li, N., Li, T., and Venkatasubramanian, S., "t-closeness: Privacy beyond k-anonymity and ℓ-diversity," *Proceedings of the IEEE 23rd International Conference on Data Engineering (ICDE)*, pp. 106–115, 2007.

[84] Li, N., Qardaji, W., and Su, D., "On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy," *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security (ASIACCS)*, pp. 32–33, 2012.

[85] Li, N., Qardaji, W., Su, D., Wu, Y., and Yang, W., "Membership privacy: A unifying framework for privacy definitions," *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security (CCS)*, pp. 889–900, 2013.

[86] Li, R. and Chang, K. C.-C., "Egonet-uiuc: A dataset for ego network research," *http://arxiv.org/abs/1309.4157*, 2013.

[87] LinkedIn *https://www.linkedin.com/*.

[88] Liu, K. and Terzi, E., "Towards identity anonymization on graphs," *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 93–106, 2008.

[89] LiveJournal *http://www.livejournal.com/*.

[90] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M., "ℓ-diversity: Privacy beyond k-anonymity," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–3, 2007.

[91] Mahmood, A. N., Kabir, M. E., and Mustafa, A. K., "New multidimensional sorting based k-anonymity microaggregation for statistical disclosure control," *Proceedings of the Security and Privacy in Communication Networks (SecureComm)*, pp. 256–272, 2012.

[92] Marti, S., Ganesan, P., and Garcia-Molina, H., "Sprout: P2p routing with social networks," *Proceedings of the Current Trends in Database Technology-EDBT 2004 Workshops*, pp. 425–435, 2005.

[93] Martin, D. J., Kifer, D., Machanavajjhala, A., and Gehrke, J., "Worst-case background knowledge for privacy-preserving data publishing," *Proceedings of the IEEE 23rd International Conference on Data Engineering (ICDE)*, pp. 126–135, 2007.

[94] McSherry, F. and Mironov, I., "Differentially private recommender systems: Building privacy into the netflix prize contenders," *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 627–636, 2009.

[95] Merener, M., "Theoretical results on de-anonymization via linkage attacks," *Transactions on Data Privacy (TDP)*, vol. 5, no. 2, pp. 377–402, 2012.

[96] Mislove, A., Marcon, M., Gummadi, K., Druschel, P., and Bhattacharjee, B., "Measurement and analysis of online social networks," *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (IMC)*, pp. 29–42, 2007.

[97] Mittal, P., Papamanthou, C., and Song, D., "Preserving link privacy in social network based systems," *Proceedings of the 20th Annual Network and Distributed System Security Symposium (NDSS)*, pp. 1–15, 2013.

[98] Mohammed, N., Chen, R., Fung, B., and Yu, P., "Differentially private data release for data mining," *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 493–501, 2011.

[99] Mohammed, N., Fung, B. C. M., Hung, P. C. K., and Lee, C., "Anonymizing healthcare data: A case study on the blood transfusion service," *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 1285–1294, 2009.

[100] Nanavati, A. A., Gurumurthy, S., Das, G., Chakraborty, D., Dasgupta, K., Mukherjea, S., and Joshi, A., "On the structural properties of massive telecom call graphs: Findings and implications," *Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM)*, pp. 435–444, 2006.

[101] Nanavati, M., Taylor, N., Aiello, W., and Warfield, A., "Herbert west - deanonymizer," *Proceedings of the 6th USENIX conference on Hot topics in security*, pp. 6–6, 2011.

[102] Narayanan, A., Shi, E., and Rubinstein, B., "Link prediction by de-anonymization: How we won the kaggle social network challenge," *Proceedings of the The 2011 International Joint Conference on Neural Networks (IJCNN)*, pp. 1825–1834, 2011.

[103] Narayanan, A. and Shmatikov, V., "Robust de-anonymization of large sparse datasets," *Proceedings of the IEEE Symposium on Security and Privacy*, pp. 111–125, 2008.

[104] Narayanan, A. and Shmatikov, V., "De-anonymizing social networks," *Proceedings of the 30th IEEE Symposium on Security and Privacy*, pp. 173–187, 2009.

[105] Nergiz, M. E., Atzori, M., and Clifton, C., "Hiding the presence of individuals from shared databases," *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 665–676, 2007.

[106] Newman, M. E. J., "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[107] Newman, M. E. J., "Networks: An introduction," *Oxford University Press*, 2010.

[108] Nilizadeh, S., Kapadia, A., and Ahn, Y.-Y., "Community-enhanced de-anonymization of online social networks," *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 537–547, 2014.

[109] Onnela, J.-P., Saramaki, J., Hyvonen, J., Szabo, G., Lazer, D., Kaski, K., Kertesz, J., and Barabasi, A.-L., "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.

[110] Orkut *https://orkut.google.com/*.

[111] Park, H. and Shim, K., "Approximate algorithms for k-anonymity," *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 67–78, 2007.

[112] Pedarsani, P., Figueiredo, D. R., and Grossglauser, M., "A bayesian method for matching two similar graphs without seeds," *Proceedings of the Allerton Conference on Communication, Control, and Computing*, pp. 1598–1607, 2013.

[113] Pedarsani, P. and Grossglauser, M., "On the privacy of anonymized networks," *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 1235–1243, 2011.

[114] Perez, S., "Twitter partners with ibm to bring social data to the enterprise," *http://techcrunch.com/2014/10/29/twitter-partners-with-ibm-to-bring-social-data-to-the-enterprise/, 2014*.

[115] Pham, H., Shahabi, C., and Liu, Y., "Ebm- an entropy-based model to infer social strength from spatiotemporal data," *Proceedings of the 2013 international conference on Management of data*, pp. 265–276, 2013.

[116] Pokec *http://pokec.azet.sk/*.

[117] Proserpio, D., Goldberg, S., and McSherry, F., "A workflow for differentially-private graph synthesis," *Proceedings of the 2012 ACM workshop on Workshop on online social networks*, pp. 13–18, 2012.

[118] Proserpio, D., Goldberg, S., and McSherry, F., "Calibrating data to sensitivity in private data analysis," *Proceedings of the 40th International Conference on Very Large Data Base (VLDB)*, vol. 7, no. 8, pp. 637–648, 2014.

[119] Qardaji, W., Yang, W., and Li, N., "Priview: Practical differentially private release of marginal contingency tables," *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pp. 1435–1446, 2014.

[120] Riordan, J., "An introduction to combinatorial analysis," *Wiley*, 1958.

[121] Ripeanu, M., Foster, I., and Iamnitchi, A., "Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design," *IEEE Internet Computing Journal*, vol. 6, pp. 50–57, 2002.

[122] Sala, A., Zhao, X., Wilson, C., Zheng, H., and Zhao, B., "Sharing graphs using differentially private graph models," *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference (IMC)*, pp. 81–98, 2011.

[123] Samarati, P., "Protecting respondents' identities in microdata release," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 13, no. 6, pp. 1010–1027, 2001.

[124] Shah, C., Capra, R., and Hansen, P., "Collaborative information seeking," *Computer*, vol. 47, no. 3, pp. 22–25, 2014.

[125] Sharad, K. and Danezis, G., "De-anonymizing d4d datasets," *Proceedings of the Workshop on Hot Topics in Privacy Enhancing Technologies (PETS)*, pp. 1–17, 2013.

[126] SlashDot *http://slashdot.org/*.

[127] Srivatsa, M. and Hicks, M., "Deanonymizing mobility traces: Using social networks as a side-channel," *Proceedings of the 2012 ACM conference on Computer and communications security (CCS)*, pp. 628–637, 2012.

[128] Sweeney, L., "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (UFKS)*, vol. 10, no. 5, pp. 557–570, 2002.

[129] Task, C. and Clifton, C., "A guide to differential privacy theory in social network analysis," *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 411–417, 2012.

[130] Terrovitis, M., Mamoulis, N., and Kalnis, P., "Privacy-preserving anonymization of set-valued data," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 115–125, 2008.

[131] Thompson, B. and Yao, D., "The union-split algorithm and cluster-based anonymization of social networks," *Proceedings of the 4th International Symposium on Information, Computer, and Communications Security (ASIACCS)*, pp. 218–227, 2009.

[132] Tummarello, K., "How 'data brokers' are striking gold," *http://thehill.com/policy/technology/207809-how-data-brokers-are-striking-gold*, 2015.

[133] Twitter *https://twitter.com/twitter*.

[134] Unnikrishnan, J. and Naini, F. M., "De-anonymizing private data by matching statistics," *Proceedings of the Allerton Conference on Communication, Control, and Computing*, pp. 1616–1623, 2013.

[135] Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P., "On the evolution of user interaction in facebook," *Proceedings of the 2nd ACM workshop on Online social networks*, pp. 37–42, 2009.

[136] Wang, K. and Fung, B. C. M., "Anonymizing sequential releases," *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 414–423, 2006.

[137] Wang, Y. and Wu, X., "Preserving differential privacy in degree-correlation based graph generation," *Transactions on data privacy (TDP)*, vol. 6, no. 2, pp. 127–145, 2013.

[138] Watts, D. J. and Strogatz, S. H., "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[139] Wills, G. J., "Nicheworks - interactive visualization of very large graphs," *Journal of Computational and Graphical Statistics*, vol. 8, no. 2, pp. 190–212, 1999.

[140] Wondracek, G., Holz, T., Kirda, E., and Kruegel, C., "A practical attack to de-anonymize social network users," *Proceedings of the 2010 IEEE Symposium on Security and Privacy*, pp. 223–238, 2010.

[141] Wong, R., Li, J., Fu, A., and Wang, K., "(α, k)-anonymity: An enhanced k-anonymity model for privacy-preserving data publishing," *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 754–759, 2006.

[142] Xiao, Q., Chen, R., and Tan, K., "Differentially private network data release via structural inference," *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 911–920, 2014.

[143] Xiao, X. and Tao, Y., "Anatomy: Simple and effective privacy preservation," *Proceedings of the 32nd international conference on Very large data bases (VLDB)*, pp. 139–150, 2006.

[144] Xiao, X. and Tao, Y., "m-invariance: Towards privacy preserving republication of dynamic datasets," *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pp. 689–700, 2007.

[145] Xiao, X. and Tao, Y., "Dynamic anonymization: Accurate statistical analysis with privacy preservation," *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 107–120, 2008.

[146] Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., and Fu, A., "Utility-based anonymization using local recording," *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 785–790, 2006.

[147] Xu, Y., Wang, K., Fu, A., and Yu, P., "Anonymizing transaction databases for publication," *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pp. 767–775, 2008.

[148] Xu, Z., Ramanathan, J., and Ramnath, R., "Identifying knowledge brokers and their role in enterprise research through social media," *Computer*, vol. 47, no. 3, pp. 26–31, 2014.

[149] Xue, M., Karras, P., Raïssi, C., Vaidya, J., and Tan, K., "Anonymizing set-valued data by nonreciprocal recording," *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1050–1058, 2012.

[150] Yang, J. and Leskovec, J., "Overlapping community detection at scale: A nonnegative matrix factorization," *Proceedings of the sixth ACM international conference on Web search and data mining (WSDM)*, pp. 587–596, 2013.

[151] Yartseva, L. and Grossglauser, M., "On the performance of percolation graph matching," *Proceedings of the first ACM conference on Online social networks*, pp. 119–130, 2013.

[152] Ying, X. and Wu, X., "Randomizing social networks: a spectrum preserving approach," *Proceedings of the SIAM International Conference on Data Mining (SDM)*, vol. 8, pp. 739–750, 2008.

[153] Youtube *https://www.youtube.com/*.

[154] Yu, H., Gibbons, P. B., Kaminsky, M., and Xiao, F., "Sybillimit: A near-optimal social network defense against sybil attacks," *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, pp. 3–17, 2008.

[155] Yu, H., Shi, C., Kaminsky, M., Gibbons, P. B., and Xiao, F., "Dsybil: Optimal sybil-resistance for recommendation systems," *Proceedings of the 30th IEEE Symposium on Security and Privacy (S&P)*, pp. 283–298, 2009.

[156] Yuan, M., Chen, L., and Yu, P., "Personalized privacy protection in social networks," *Proceedings of the VLDB Endowment*, vol. 4, no. 2, pp. 141–150, 2010.

[157] Zhang, Q., Koudas, N., Srivastava, D., and Yu, T., "Aggregate query answering on anonymized tables," *Proceedings of the IEEE 23rd International Conference on Data Engineering (ICDE)*, pp. 116–125, 2007.

[158] Zhou, B. and Pei, J., "Preserving privacy in social networks against neighborhood attacks," *Proceedings of the IEEE 24th International Conference on Data Engineering (ICDE)*, pp. 506–515, 2008.

[159] Zhou, B., Pei, J., and Luk, W.-S., "A brief survey on anonymization techniques for privacy preserving publishing of social network data," *ACM SIGKDD Explorations Newsletter*, vol. 10, no. 2, pp. 12–22, 2008.

[160] Zou, L., Chen, L., and Özsu, M. T., "K-automorphism: A general framework for privacy preserving network publication," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 946–957, 2009.